

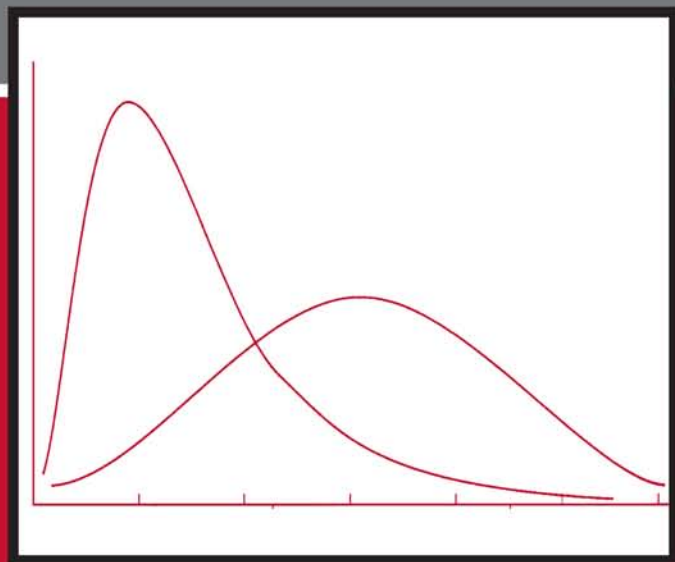
Data CD
Enclosed

SAS

A Modern Approach

INTERMEDIATE STATISTICS

THIRD EDITION



SPSS

JAMES P. STEVENS

INTERMEDIATE STATISTICS

A Modern Approach

THIRD EDITION

INTERMEDIATE STATISTICS

A Modern Approach

THIRD EDITION

James P. Stevens



Lawrence Erlbaum Associates
Taylor & Francis Group

New York London

Cover design by Kathryn Houghtaling.

Lawrence Erlbaum Associates
Taylor & Francis Group
270 Madison Avenue
New York, NY 10016

Lawrence Erlbaum Associates
Taylor & Francis Group
2 Park Square
Milton Park, Abingdon
Oxon OX14 4RN

© 2007 by Taylor & Francis Group, LLC

Lawrence Erlbaum Associates is an imprint of Taylor & Francis Group, an Informa business

Printed in the United States of America on acid-free paper

10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8058-5466-4 (Softcover) 978-0-8058-5465-7 (Hardcover)

No part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
<http://www.taylorandfrancis.com>

Contents

<i>Preface</i>	xi
<u>Chapter 1</u>	
<i>INTRODUCTION</i>	1
1.1 Focus and Overview of Topics	
1.2 Some Basic Descriptive Statistics	
1.3 Summation Notation	
1.4 t Test for Independent Samples	
1.5 t Test for Dependent Samples	
1.6 Outliers	
1.7 SPSS and SAS Statistical Packages	
1.8 SPSS for Windows—Release 12.0	
1.9 Data Files	
1.10 Data Entry	
1.11 Editing a Dataset	
1.12 Splitting and Merging Files	
1.13 Two Ways of Running Analyses on SPSS	
1.14 SPSS Output Navigator	
1.15 SAS and SPSS Output for Correlations, Descriptives, and t Tests	
1.16 Data Sets on Compact Disk	
Appendix Obtaining the Mean and Variance on the TI-30Xa Calculator	
<u>Chapter 2</u>	
<i>ONE WAY ANALYSIS OF VARIANCE</i>	45
2.1 Introduction	
2.2 Rationale for ANOVA	
2.3 Numerical Example	
2.4 Expected Mean Squares	
2.5 MS_w and MS_b as Variances	

2.6	A Linear Model for the Data
2.7	Assumptions in ANOVA
2.8	The Independence Assumption
2.9	ANOVA on SPSS and SAS
2.10	Post Hoc Procedures
2.11	Tukey Procedure
2.12	The Scheffé Procedure
2.13	Heterogeneous Variances and Unequal Group Sizes
2.14	Measures of Association (Variance Accounted For)
2.15	Planned Comparisons
2.16	Test Statistic for Planned Comparisons
2.17	Planned Comparisons on SPSS and SAS
2.18	The Effect of an Outlier on an ANOVA
2.19	Multivariate Analysis of Variance
2.20	Summary
	Appendix

Chapter 3

POWER ANALYSIS

105

3.1	Introduction
3.2	t Test for Independent Samples
3.3	A Priori and Post Hoc Estimation of Power
3.4	Estimation of Power for One Way Analysis of Variance
3.5	A Priori Estimation of Subjects Needed for a Given Power
3.6	Ways of Improving Power
3.7	Power Estimation on SPSS MANOVA
3.8	Summary

Chapter 4

FACTORIAL ANALYSIS OF VARIANCE

123

4.1	Introduction
4.2	Numerical Calculations for Two Way ANOVA
4.3	Balanced and Unbalanced Designs
4.4	Higher Order Designs
4.5	A Comprehensive Computer Example Using Real Data
4.6	Power Analysis
4.7	Fixed and Random Factors
4.8	Summary
	Appendix Doing a Balanced Two Way ANOVA With a Calculator

Chapter 5*REPEATED MEASURES ANALYSIS*

181

- 5.1 Introduction
- 5.2 Advantages and Disadvantages of Repeated Measures Designs
- 5.3 Single Group Repeated Measures
- 5.4 Completely Randomized Design
- 5.5 Univariate Repeated Measures Analysis
- 5.6 Assumptions in Repeated Measures Analysis
- 5.7 Should We Use the Univariate or Multivariate Approach?
- 5.8 Computer Analysis on SAS and SPSS for Example
- 5.9 Post Hoc Procedures in Repeated Measures Analysis
- 5.10 One Between and One Within Factor—A Trend Analysis
- 5.11 Post Hoc Procedures for the One Between and One Within Design
- 5.12 One Between and Two Within Factors
- 5.13 Totally Within Designs
- 5.14 Planned Comparisons in Repeated Measures Designs
- 5.15 Summary

Chapter 6*SIMPLE AND MULTIPLE REGRESSION*

219

- 6.1 Simple Regression
- 6.2 Assumptions for the Errors
- 6.3 Influential Data Points
- 6.4 Multiple Regression
- 6.5 Breakdown of Sum of Squares in Regression and F Test for Multiple Correlation
- 6.6 Relationship of Simple Correlations to Multiple Correlation
- 6.7 Multicollinearity
- 6.8 Model Selection
- 6.9 Two Computer Examples
- 6.10 Checking Assumptions for the Regression Model
- 6.11 Model Validation
- 6.12 Importance of the Order of Predictors in Regression Analysis
- 6.13 Other Important Issues
- 6.14 Outliers and Influential Data Points
- 6.15 Further Discussion of the Two Computer Examples
- 6.16 Sample Size Determination for a Reliable Prediction Equation
- 6.17 ANOVA as a Special Case of Regression Analysis
- 6.18 Summary of Important Points
- Appendix The PRESS Statistic

Chapter 7*ANALYSIS OF COVARIANCE*

285

- 7.1 Introduction
- 7.2 Purposes of Covariance
- 7.3 Adjustment of Posttest Means
- 7.4 Reduction of Error Variance
- 7.5 Choice of Covariates
- 7.6 Numerical Example
- 7.7 Assumptions in Analysis of Covariance
- 7.8 Use of ANCOVA with Intact Groups
- 7.9 Computer Example for ANCOVA
- 7.10 Alternative Analyses
- 7.11 An Alternative to the Johnson–Neyman Technique
- 7.12 Use of Several Covariates
- 7.13 Computer Example with Two Covariates
- 7.14 Summary

Chapter 8*HIERARCHICAL LINEAR MODELING*

321

- 8.1 Introduction
- 8.2 Problems Using Single-Level Analyses of Multilevel Data
- 8.3 Formulation of the Multilevel Model
- 8.4 Two-Level Model—General Formulation
- 8.5 HLM6 Software
- 8.6 Two Level Example—Student and Classroom Data
- 8.7 HLM Software Output
- 8.8 Adding Level One Predictors to the HLM
- 8.9 Addition of a Level Two Predictor to a Two Level HLM
- 8.10 Evaluating the Efficacy of a Treatment
- 8.11 Final Comments on Hlm

Appendix A*DATA SETS*

365

- A.1 Clinical Data
- A.2 Alcoholics Data
- A.3 Sesame Street Data
- A.4 Headache Data
- A.5 Cartoon Data

A.6 Attitude Data
A.7 National Academy of Sciences Data
A.8 Agresti Home Sales Data

Appendix B
STATISTICAL TABLES 399

Table B.1 Critical Values for F
Table B.2 Percentile Points of Studentized Range Statistic
Table B.3 Critical Values for Dunnett's Test
Table B.4 Critical Values for F (max) Statistic
Table B.5 Critical Values for Bryant-Paulson Procedure

Appendix C
POWER TABLES 413

Table C.1 Power of F Test at $\alpha = .05, u = 1$
Table C.2 Power of F Test at $\alpha = .05, u = 2$
Table C.3 Power of F Test at $\alpha = .05, u = 3$
Table C.4 Power of F Test at $\alpha = .05, u = 4$
Table C.5 Power of F Test at $\alpha = .10, u = 1$
Table C.6 Power of F Test at $\alpha = .10, u = 2$
Table C.7 Power of F Test at $\alpha = .10, u = 3$
Table C.8 Power of F Test at $\alpha = .10, u = 4$

References 423

Answers to Selected Exercises 431

Author Index 453

Subject Index 457

Preface

This book is written for behavioral and social science students at the advanced undergraduate or beginning graduate level. The text emphasizes conceptual understanding, the effective use of statistical software to run the analyses, and the correct interpretation of results. Two statistical software packages, SAS and SPSS, are an integral part of each chapter. An annotated printout is given from at least one of the programs for each analysis. The annotations highlight what the numbers mean and how to interpret the results. The explanation appears on the printout or on the same page to enhance learning efficiency. The assumptions underlying each analysis are given special attention, and the reader is shown how to test the critical assumption(s) using SAS and SPSS. Power analysis is an integral part of the book. There are no computational formulas in this text. I took the position that they were not needed many years ago, and it is even truer today.

The instructional mix of strategies that is employed to illustrate each statistical technique consists of two parts (a) First, I use definitional formulas on small data sets to convey conceptual insight into what is being measured, and (b) Then, I proceed directly to the packages to efficiently process data. I feel very strongly about using these strategies.

The most significant change in this edition is the addition of a chapter on hierarchical linear modeling using HLM6. This material is important because correlated observations occur frequently in social science research and just a SMALL amount of dependence causes the type I error rate to be several times greater than one wishes! Since HLM involves a series of regressions, this new chapter is placed after the material on regression. The distinction between fixed and random factors is important, and so it is emphasized. The chapter on HLM was written by Dr. Natasha Beretvas of the University of Texas at Austin. I thank her very much for her contribution.

The third edition features newer versions of SPSS (Release 12.0) and SAS (Release 8.0). Much of the material on importing data into SAS or SPSS that previously appeared in chapter 1 was deleted. Importing data into these two programs is now much easier so this material was no longer necessary.

The exercises involve a mixture of numerical, conceptual and computer related problems. I have de-emphasized purely numerical exercises, for I agree entirely with Cobb (1987, p. 323) that, “computing rules are just the skin of our subject; it is focus that reveals the skeleton of fundamental concepts and connections that hold

the body of knowledge together.” Regarding exercises, it is important to note that there are 3 new exercises for each chapter. Answers are provided for half of the exercises and an *Instructor’s Solutions CD* is available to adopters. A computer example of real data integrates many of the concepts. A CD containing all of the book’s data sets is included in the back of the book.

The reader should have a background of a one quarter course in statistics that covered at least the t tests for independent and dependent samples.

I am very grateful to the reviewers of this text: Dale Berger of Claremont Graduate University, Michael Milburn of University of Massachusetts, Mary Lou Kerwin of Rowan University, Gordon Brooks of Ohio University, and Roderick Gillis of the University of Miami. I am also indebted to some individuals at my publisher. Larry Erlbaum continues to be very supportive. Debra Riegert was instrumental in motivating me to write this third edition.

Jim Stevens

Introduction

CONTENTS

- 1.1 Focus and Overview of Topics
- 1.2 Some Basic Descriptive Statistics
- 1.3 Summation Notation
- 1.4 t Test for Independent Samples
- 1.5 t Test for Dependent Samples
- 1.6 Outliers
- 1.7 SPSS and SAS Statistical Packages
- 1.8 SPSS for Windows—Release 12.0
- 1.9 Data Files
- 1.10 Data Entry
- 1.11 Editing a Dataset
- 1.12 Splitting and Merging Files
- 1.13 Two Ways of Running Analyses on SPSS
- 1.14 SPSS Output Navigator
- 1.15 SAS and SPSS Output for Correlations, Descriptives, and t Tests
- 1.16 Data Sets on Compact Disk
- Appendix Obtaining the Mean and Variance on the TI-30Xa Calculator

1.1 FOCUS AND OVERVIEW OF TOPICS

This book has been written for applied social science researchers at the advanced undergraduate or beginning graduate level. It is assumed that you have had a one quarter course in beginning statistics that covered measures of central tendency, measures of variability, standard scores (z , T , stanines, etc.), correlation, and inferential statistics, including at least the t tests for independent and dependent samples. In the next four sections of this chapter, we review briefly some descriptive statistics, summation notation, and testing for a “significant” difference. These

sections are not intended to thoroughly teach this material again, but to refresh your memory.

The emphasis in the book is on conceptual understanding of the statistical techniques, learning how to effectively use statistical software to run the analyses, and learning how to interpret the computer printout that results from such runs. The two major statistical packages, SAS (Statistical Analysis System) and SPSS (Statistical Package for the Social Sciences), are an integral part of this book. Details on SAS and SPSS are given in Section 1.7. I have attempted to make the text as practical as possible. To accent the practical emphasis, nine real data sets have been provided in Appendix A in the back of the book. For convenience, these data sets are also available on a CD. Some of the exercises in the chapters involve running these data sets, or a part of a real data set. Singer and Willett (1988) have provided an excellent annotated bibliography, indicating where numerous other real data sets may be found.

The instructional mix of strategies adopted to illustrate each statistical technique involves two parts:

1. First, we illustrate each technique using *definitional* formulas on small data sets. These formulas are useful in yielding conceptual insight into what is being measured or quantified. As a simple example, the definitional formula for sample variance is

$$s^2 = [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] / (n - 1)$$

This formula shows very clearly that variance is measuring how much the scores for the subjects scatter or disperse about the mean.

2. Then we move directly to the computer, that is, to the statistical packages, to show how to efficiently process data. And more importantly, how to interpret the printout from the packages. In practice, analyses will very likely be run on one or more of these packages, and thus it is important to become familiar with them.

Now we give an overview of the topics in the book. The reader may recall that the *t* test for independent samples is appropriate for comparing two groups to determine whether they differ on the average on a dependent variable. But what if we wish to compare more than two groups *simultaneously* on a dependent variable? For example, we wish to compare the effect of four counseling methods on attitude toward education. Then a statistical procedure called analysis of variance is needed. This technique is covered in Chapter 2. Suppose that for this example there was reason to believe that the sex of the subjects might moderate the effect of the counseling methods, and we wanted to check this possibility. This would lead us to a more complicated analysis of variance design, since we are examining the effect of two independent variables (sex and counsel-

ing method) on attitude toward education. It is an example of a factorial design. These designs are covered in Chapter 4.

Chapter 3 deals with power analysis. The power of a statistical test is the probability of rejecting the null hypothesis when it is false. Although it may seem obvious that we would want to achieve this, many researchers in the literature have failed to do so, as Cohen (1969) and others have pointed out. The reason is that power is generally inadequate with small group sizes (especially with 20 or less subjects per group), and in some areas of research such sample sizes are quite common for pragmatic or other reasons. Chapter 3 provides a detailed and practical approach to estimating the power of completed studies and also for estimating sample size required for adequate power in an upcoming study.

In Chapter 5 we treat the class of situations in which the same subjects are measured more than twice on a dependent variable. For example, suppose a dietitian wishes to assess the immediate and long term effects of a behavior modification approach on weight loss for a group of overweight men. She measures the weight loss immediately following treatment and then 7 additional times (in three month intervals) over a two year period. The appropriate statistical analysis here is a different type of analysis of variance from that in Chapter 2, called repeated measures analysis. The simplest case of a repeated measures design measures the subjects just twice, e.g., pretest—treatment—posttest. The investigator is interested in testing for a significant gain or change on the dependent variable, and the appropriate test is the *t* test for correlated (dependent) samples that you studied in beginning statistics.

Chapter 6, which is a new addition to my intermediate text, deals with multiple regression. Much of the material is taken from my multivariate text (Stevens, 1996). Multiple regression is a much used and abused technique. One of the problems is that many researchers use multiple regression without validating their results on an independent sample of data. I have made validating the model a major theme in this chapter.

Analysis of covariance is now found in Chapter 7. This technique combines analysis of variance and regression analysis. Because of this, and because of the suggestions of two reviewers of this edition, I have covariance after regression and ANOVA. A covariate is a variable that is significantly correlated with the dependent variable. Analysis of covariance can be quite helpful in randomized studies, that is, studies where the subjects have been randomly assigned to the treatments, in increasing the sensitivity (power) of an experiment.

Analysis of variance procedures and multiple regression are used very often in the literature. Thus it is important to learn this material in order to be able to intelligently and critically read the literature.

1.2 SOME BASIC DESCRIPTIVE STATISTICS

The measure of central tendency that is used most frequently is the mean or average for a set of scores. It is defined as

$$\bar{x} = (x_1 + x_2 + \cdots + x_n) / n$$

where n is the number of subjects and x_1 , is the score for subject 1 on variable x , x_2 is the score for subject 2, etc. The mean is an example of a summary statistic—it summarizes an important or salient feature of a set of data. For example, if you are told that the average weight for a pro football lineman is 280 pounds, or that the average income of people living in a certain community is \$80,000, each of these numbers packs a message. The average weight of 280 pounds indicates that the weights of most linemen tend to cluster around that value, and the income of \$80,000 means that the incomes of most people in that community cluster around \$80,000. These statements are accurate provided that there are no extreme values or outliers (see Section 1.6).

Although the mean is useful in characterizing one important feature of a set of data, it can be misleading just by itself. To see this consider the following scores for three groups of 10 children each on a 20 item pretest in mathematics:

<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>
10	10	10
13	11	18
7	11	2
12	10	13
13	12	17
11	11	3
8	11	8
14	10	15
9	12	19
12	11	4
$\bar{x}_1 = 10.9$	$\bar{x}_2 = 10.9$	$\bar{x}_3 = 10.9$

On the average there is no difference between these three groups of children. However, there *is* a major difference among the groups in terms of variability of the scores about the mean. One can see intuitively that there is the least variability for group 2 (since the scores cluster very tightly about the mean of 10.9), while variability is greatest for group 3. This differential amount of variability would have definite instructional implications, if you had to teach one of these three groups mathematics. Other things being equal, group 2 would be easier to teach since they are all at about the same level of ability.

To quantify the amount of variability in a set of scores we use the sample variance s^2 , the *definitional* formula of which is

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n - 1}$$

Notice that variance simply measures how much the scores vary about the mean. Now we find the variances for the three groups of children. Although the emphasis in this book is on using the computer for doing statistical analysis, there is a wide array of very inexpensive calculators that are conveniently used for calculating the mean and variance for a set of data. In Appendix 1 at the end of this chapter we give the details for the TI-30Xa for the children in group 1. The variances for the three groups are: $s_1^2 = 5.43$, $s_2^2 = .54$, and $s_3^2 = 41.43$.

Summary statistics like the mean and variance are especially useful in comparing different data sets (groups of subjects) on the same variable. Consider the following two sets of scores, which represent the age of 25 automobile salesmen in the United States and 25 automobile salesmen in Western Europe:

<i>United States</i>					<i>Western Europe</i>				
23	63	25	22	32	43	26	30	27	40
56	30	34	56	30	35	48	36	47	41
25	48	44	27	26	34	45	30	38	33
38	26	30	39	30	35	44	24	33	40
36	32	36	38	33	31	23	29	37	28

It is far from obvious by just looking at these sets how the ages for the two groups differ, if at all. Computation of the mean and variance for the groups yields: U.S. ($\bar{x} = 35.16$, $s^2 = 117.22$) and Western Europe ($\bar{x} = 35.08$, $s^2 = 51.16$). These statistics indicate that the average age is about the same for the groups and that the variability in age for the U.S. salesmen is over twice that for the Western Europeans.

1.3 SUMMATION NOTATION

The reader probably was exposed to the summation operator in an introductory statistics course. Nevertheless, a brief review of some basic properties of Σ (sigma) will be helpful. The symbol Σ means “take the sum of.” Suppose we had measured 50 subjects on anxiety. The sum of their scores is

$$x_1 + x_2 + x_3 \dots + x_{50}$$

This sum can be expressed concisely using Σ as follows:

$$\sum_{i=1}^{50} x_i$$

The first term (x_1) is obtained by setting $i = 1$, the second term (x_2) by setting $i = 2$, on down to the last term (x_{50}) for $i = 50$. The quantity i is called the index of summation; it is what we are summing on. Let us consider a few more examples to illustrate. Suppose we have measured 75 subjects on a variable y and wish to represent the sum of those scores using Σ . Then it would look like this:

$$y_1 + y_2 + \cdots + y_{75} = \sum_{i=1}^{75} y_i$$

Or if we had 100 subjects measured on variable z and wish the sum of the scores for subjects 3 through 100, then we have

$$z_3 + z_4 + \cdots + z_{100} = \sum_{i=3}^{100} z_i$$

If the limits are understood, then they are dropped, and we would just write Σz_i . Note that the mean for a set of n scores can be written using Σ :

$$\bar{x} = (x_1 + x_2 + \cdots + x_n) / n = \Sigma x_i / n$$

Often we may wish to concisely represent a sum of squares of some type. Suppose we have n subject scores (x_1, x_2, \dots, x_n) and wish to denote the sum of the squared scores. This is

$$x_1^2 + x_2^2 + \cdots + x_n^2 = \Sigma x_i^2$$

The sample variance for a set of scores involves a sum of squares (squared deviations):

$$\begin{aligned} s^2 &= [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] / (n - 1) \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 / (n - 1) \end{aligned}$$

or as $s^2 = \Sigma (x - \bar{x})^2 / (n - 1)$ if the limits are understood.

Example

Evaluate Σx_i^2 , where $x_1 = 10$, $x_2 = 8$, $x_3 = 13$, and $x_4 = 5$.

$$\sum x_i^2 = x_1^2 + x_2^2 + x_3^2 + x_4^2 = 10^2 + 8^2 + 13^2 + 5^2 = 358$$

The following four properties of the summation operator are useful to know:

1.
$$\begin{array}{ccc} \sum (x + y) & = & \sum x + \sum y \\ \text{summation} & & \text{sum} \\ \text{of sum} & & \text{of summations} \end{array}$$
2.
$$\begin{array}{ccc} \sum (x - y) & = & \sum x - \sum y \\ \text{summation of} & & \text{difference} \\ \text{difference} & & \text{in summations} \end{array}$$
3.
$$\sum cx_i = c \sum x_i \text{ (a constant } c \text{ can be moved across the summation)}$$

To show that property 3 holds note that

$$\sum cx_i = cx_1 + cx_2 + \cdots + cx_n = c(x_1 + x_2 + \cdots + x_n) = c \sum x_i$$

4.
$$\sum c = nc \text{ (summing over } n \text{ subjects)}$$

The constant c mentioned in properties 3 and 4 can appear in many different subtle ways. To illustrate that and also to show how to apply several of the above properties, we will prove that the mean of a set of z scores is 0.

Denote the z scores by z_1, z_2, \dots, z_n . Then by definition of a mean we have

$$\bar{z} = \sum z_i / n$$

To show that $\bar{z} = 0$ it suffices to show that $\sum z_i = 0$.

By definition $z_i = (x_i - \bar{x}) / s$. Therefore by substitution:

$$\sum z_i = \sum (x_i - \bar{x}) / s$$

Note that $1/s$ is a constant here; that is, it does not depend on i (index of summation). By property 3 we can move it across the summation and write

$$\sum z_i = (1/s) \sum (x_i - \bar{x})$$

Now by property 2 we can further rewrite this as

$$\sum z_i = (1/s) [\sum x_i - \sum \bar{x}]$$

Next, \bar{x} is a constant and thus by property 4 we have that $\sum \bar{x} = n\bar{x}$. Also, since $\bar{x} = \sum x_i / n$ (by definition), this implies that $\sum x = n\bar{x}$. Plugging these values in we obtain

$$\sum z_i = (1/s)[n\bar{x} - n\bar{x}] = (1/s) \cdot 0 = 0$$

1.4 t TEST FOR INDEPENDENT SAMPLES

As an example we consider a study by air force psychologists conducting research into the relative effectiveness of training pilots. The first method makes use of computer simulated flight while the second uses traditional flight instruction. The 18 subjects were randomly assigned to the two methods and the following performance test scores were obtained:

<i>Computer Simulation</i>	<i>Flight</i>
2	1
5	1
5	2
6	3
6	3
7	4
8	5
9	7
	7
	8

We wish to test at the $\alpha = .05$ level of significance whether the average performance for the two groups is different. Recall that we wish to test the null hypothesis (H_0) that the population means are equal:

$$H_0: \mu_1 = \mu_2$$

It is called the null hypothesis because saying the population means are equal is equivalent to saying that the difference in the means is 0, i.e., $\mu_1 - \mu_2 = 0$, or that the difference is null.

Remember that level of significance is our probability of making a type I error. *Type I error is the probability of rejecting the null hypothesis when it is true, or saying the groups differ when they don't.* This type of error can not be eliminated; however, we can and do control the risk by setting $\alpha = .05$ or $.01$. Then there is only a 5% or 1% chance of making this type of error.

It should be recalled that the *t* test is based on the following three assumptions:

1. Normality—the scores on the dependent variable are normally distributed in each group.
2. Homogeneity of variance—the population variances are equal for the two groups.
3. Independence of the observations—each subject's score on the dependent variable is not affected by other subjects in the same treatment group.

Briefly, considerable research has shown that a violation of the normality assumption is of little consequence. Unequal variances will distort the type I error rate appreciably *only if* the group sizes are sharply unequal (largest/smallest > 1.5). Finally, dependent observations have a very serious effect on type I error rate. We discuss violations of assumptions in considerable detail in Chapter 2.

To test H_0 we use the following t statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}, \text{ with } (n_1 + n_2 - 2)df \quad (1)$$

where $s_p^2 = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$ is the pooled estimate of the assumed common population variance for the groups (the homogeneity of variance assumption). Now, s_1^2 and s_2^2 are the sample variances for groups 1 and 2, while n_1 and n_2 are the respective group sizes. This test statistic can be calculated relatively easily by obtaining the mean and variance for each group with the TI-30Xa or some other calculator. With the variances obtained, we find that

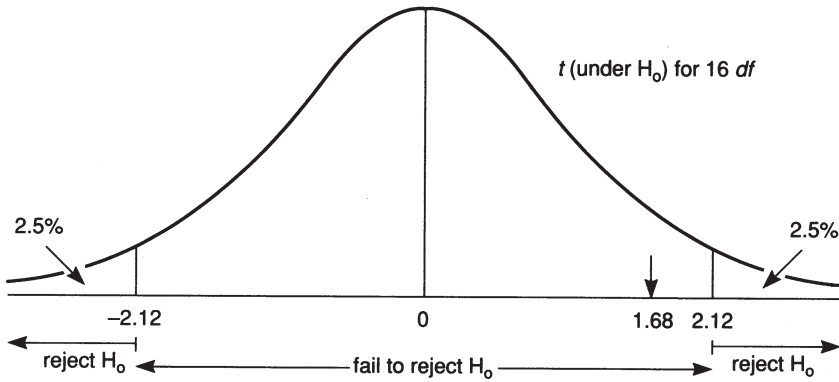
$$s_p^2 = [(8 - 1)4.57 + (10 - 1)6.54]/(10 + 8 - 2) = 5.68$$

Using Equation 1 we calculate

$$t = \frac{6 - 4.1}{\sqrt{5.68(1/8 + 1/10)}} = \frac{1.9}{1.13} = 1.68$$

Recall that we decided to reject H_0 only if the value of t obtained was *very unlikely* (would occur only 5% of the time) under the assumption of equal population means. The sampling distribution of t values (under the null hypothesis of equal population means) for this case is shown on the following page.

From the figure we can see that only 2.5% of the time will we obtain a t value greater than 2.12 or less than -2.12 *if the null hypothesis is true*. The 2.12 and -2.12 are called critical values because they are critical or pivotal points for our decision on H_0 . Note that the critical values define the critical regions, where rejection of H_0 occurs. In general, if the value of t is greater than (in absolute value) the critical value, we will reject H_0 ; otherwise, we fail to reject. In this case since $t = 1.68$ is not in the critical region, we fail to reject H_0 .



The null hypothesis could have also been tested using a *confidence interval*. Confidence intervals are an important part of inferential statistics. The confidence interval will give us a range of values within which the population mean difference lies with a certain probability (or confidence). For the above t test for independent samples the confidence interval is given by:

$$(\bar{x}_1 - \bar{x}_2) - t_{.05; df} s_{\bar{x}_1 - \bar{x}_2} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{.05; df} s_{\bar{x}_1 - \bar{x}_2}$$

where $t_{.05; df}$ denotes the two tailed critical value at .05 with $(n_1 + n_2 - 2)$ degrees of freedom and $s_{\bar{x}_1 - \bar{x}_2}$ is the denominator of the t statistic. Thus, the 95% confidence interval for the above problem is given by

$$1.9 - 2.12 (1.13) < \mu_1 - \mu_2 < 1.9 + 2.12 (1.13) \\ -.496 < \mu_1 - \mu_2 < 4.296$$

Since this interval covers (crosses) 0, this means 0 is a possible value for $\mu_1 - \mu_2$, which means it is likely that $\mu_1 - \mu_2 = 0$ or that $\mu_1 = \mu_2$. Since it is possible that the population means are equal we would not reject the null hypothesis. On the other hand, if the confidence interval does *not* cross 0 then we conclude there is a significant difference between the groups, because this would mean 0 is not a possible value for the population mean difference. *Confidence intervals are more informative than a test of significance because they not only test the null hypothesis but also give us a range of values that is useful in judging the practical significance of results.* We discuss the practical significance of results more in Chapter 2 on one way analysis of variance.

1.5 *t* TEST FOR DEPENDENT SAMPLES

The *t* test for dependent samples is appropriate in a variety of situations, of which the following three are common:

- a. Pretest–treatment–posttest.
- b. Two groups of matched or paired subjects are compared on some dependent variable. For example, 16 girl beginners are matched on SES, I.Q., number of children in the family, and general health. Eight of the girls had attended kindergarten; the other 8 had not. We wish to determine whether they differ on a test of first grade readiness.
- c. We are comparing naturally occurring correlated pairs, such as twins, husband and wife, parent and child, etc.

Our numerical example does *not* fit into the above three categories.

Example

A political candidate wishes to determine if endorsing increased social spending is likely to affect her standing in the polls. She has access to data on the popularity of several other candidates who have endorsed social spending. The data were available both before and after the candidates announced their positions on the issue, as follows:

<i>Candidate</i>	Popularity		<i>d_i</i>
	<i>Before</i>	<i>After</i>	
1	42	43	1
2	41	45	4
3	50	56	6
4	52	54	2
5	58	65	7
6	32	29	–3
7	39	46	7
8	42	48	6
9	48	47	–1
10	47	53	6

The *d_i* are the difference scores in popularity and are fundamental in defining the test statistic for correlated samples:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}, \text{ with } (n-1) \text{ } df \quad (2)$$

where \bar{d} is the average difference score, s_d is the standard deviation for the difference scores and n is the number of subjects or matched pairs. By using the TI-30Xa calculator on the difference scores above one obtains $\bar{d} = 3.5$ and $s_d = 3.57$.

Therefore, t is calculated as $t = \frac{3.5}{3.57/\sqrt{10}} = 3.097$. The critical value at the

.05 level is $t_{.05;9} = 2.262$. Since the value of the test statistic is greater than the critical value, we reject and conclude that mean popularity after endorsement is greater than the mean popularity before endorsement.

Note that the mean difference is equal to the difference in the means, as we show below, where x_a and x_b denote the scores after and before:

$$\bar{d} = \sum d_i / n = \sum (x_a - x_b) / n = \sum x_a / n - \sum x_b / n = \bar{x}_a - \bar{x}_b$$

If we rewrite the equation for the t test for independent samples as

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$$

then by placing this side by side with the t test for dependent samples we can see that they are structurally identical:

Independent t	Dependent t
$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}}$	$t = \frac{\bar{x}_a - \bar{x}_b}{s_p \sqrt{1/n}}$

The numerator in each case involves an estimate of the difference in the means; in the first case for the two groups and in the second case for the matched pairs or subjects on two different occasions. In the denominators, s_p and s_d provide estimates of the amount of sampling error for each mean difference.

1.6. OUTLIERS

An outlier is a data point which splits off or is very different from the rest of the data. Specific examples of outliers would be an I.Q. of 160 (among normal subjects), or a weight of 350 lbs. in a normal population of subjects. It is very important to detect outliers because they can have a dramatic effect on the results of any statistical analysis.

Outliers can occur because of two fundamental reasons:

1. a data recording or entry error was made, or
2. the subjects are different from the rest.

The first type of outlier can be identified by *always* listing the data and checking to make sure the data has been read in accurately. Consider the following small data set with two variables:

	X_1	X_2	$Zscore(X_1)$	$Zscore(X_2)$
1	101.00	68.00	-.25078	.53882
2	92.00	46.00	-.77566	-.94293
3	90.00	50.00	-.89230	-.67352
4	107.00	59.00	.09914	-.06735
5	98.00	50.00	-.42574	-.67352
6	150.00	66.00	2.60691	.40411
7	108.00	54.00	.15746	-.40411
8	110.00	51.00	.27410	-.60617
9	103.00	59.00	-.13414	-.06735
10	94.00	97.00	-.65902	2.49202
Total N	10	10	10	10

Do you see any outlier(s) for x_1 , or x_2 ? Subject 6 is an outlier for x_1 ; notice how 150 splits off dramatically from the rest of the scores, which fall in the range from 90 to 110. Subject 10 is an outlier for x_2 , since the score of 97 splits off sharply from the rest of the scores, which fall mostly in the range from about 50 to the mid 60s.

The z scores make the outliers quite apparent, and the z scores are very high since Shiffler (1988) has shown that the *largest possible* z score in a sample of size 10 is 2.846. We elaborate on this later on in the section.

You actually encountered the notion of an outlier in a beginning statistics course, although it may not have been called that by the instructor. In discussing measures of central tendency, your instructor probably indicated that whenever you have extreme scores in a set of data, the median should be used to characterize the data, rather than the mean. Extreme scores are called outliers here. The reason you were told to use the median is that it is essentially unaffected by extreme scores whereas the mean is drastically affected. Consider the following set of data: 2, 3, 5, 6, 44. The last number is an outlier. If we were to use the mean (12), it would be quite misleading in characterizing the data set, as there are no scores around 12. The median, on the other hand, is 5 and does indicate where most of the scores lie (although there are only 5 of them).

To show the dramatic effect an outlier can have on a correlation, consider the two scatterplots in Figure 1.1. Notice how inclusion of the outlier in each case *drastically* changes the interpretation of the results. For Case A there is no relationship without the outlier but there is a strong relationship with the outlier, while for Case B the relationship changes from strong (without outlier) to weak.

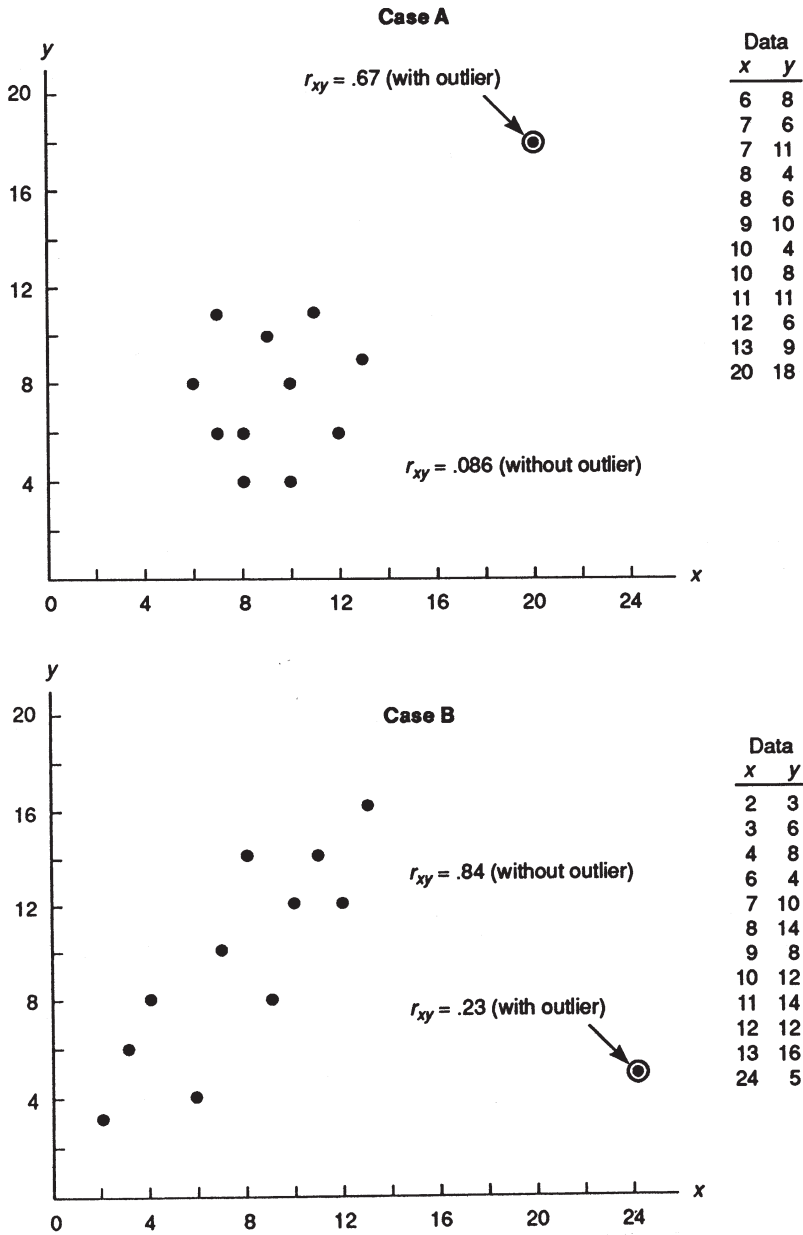


FIGURE 1.1 The Effect of an Outlier on a Correlation Coefficient.

From the above it should be clear that *it is very important to identify outliers and then decide what to do about them*. Why? Because we want our analysis results reflecting most of the data, and not being unduly influenced by just 1 or 2 errant points.

Detecting Outliers

If the variable is approximately normally distributed, then z scores around 3 in absolute value should be considered as potential outliers. Why? Because in an approximate normal distribution about 99% of the scores should lie within three standard deviations of the mean. Therefore, any z value > 3 indicates a value very unlikely to occur. Of course, if n is large (say > 100), then simply by chance we might expect a few subjects to have z scores > 3 and this should be kept in mind. However, even for *any type of distribution* the above rule is reasonable, although we might consider extending the rule to $z > 4$. It was shown many years ago that regardless of how the data are distributed the percentage of observations that are, contained within k standard deviations of the mean must be *at least* $(1 - 1/k^2) \cdot 100\%$. The above holds only for $k > 1$.

Shiffler (1988) has shown that the largest possible value z value in a data set of size n is bounded by $(n-1)\sqrt{n}$. This means for $n = 10$ the largest possible z is 2.846 and for $n = 11$ the largest possible z is 3.015. Thus, for small sample size any data point with a z around 2.5 should be seriously considered as a possible outlier.

1.7 SPSS AND SAS STATISTICAL PACKAGES

The Statistical Analysis System (SAS) and the Statistical Package for the Social Sciences (SPSS) were selected for use in this text for several reasons.

1. They are very widely distributed.
2. They are easy to use.
3. They do a very wide range of analyses—from simple descriptive statistics to various analysis of variance designs to complex multivariate analyses.
4. They are well documented, having been in development for over two decades.

The control language that is used by both packages is quite natural, and you will see that with a little practice complex analyses are run quite easily, and with a small set of control line instructions. A major change from the previous edition of this text is the advent of Windows and running analyses by simply clicking a series of buttons. It is assumed that the reader will be either running a Windows version of one or both of these packages on a desktop computer, or perhaps a notebook com-

puter, or running the analyses from the program editor (called this in SAS, or from the syntax editor, as called by SPSS). We illustrate the SPSS for Windows 12.0 in some detail. Examples are considered where the data is part of the control lines.

Structurally, an SAS program is composed of three fundamental blocks:

1. Statements setting up the data.
2. The data lines.
3. Procedure (PROC) statements—procedures are SAS computer programs which read the data and do various statistical analyses.

To illustrate how to set up the control lines, suppose we wish to compute the correlations between locus of control, achievement motivation, and achievement in language for a hypothetical set of 9 subjects. First we create a data set and give it a name. The name *must* begin with a letter and be 8 or less characters. Let us call the data set LOCUS. Now, each SAS statement *must* end with a semicolon. So our first SAS line looks like this

```
DATA LOCUS;
```

The next statement needed is called an INPUT statement. This is where we give names for our variables and indicate the format of the data (i.e., how the data is arranged on each line). We will use what is called free format. With this format the scores for each variable do not have to be in specific columns. However, at least one blank column must separate the score for each variable from the next variable. Furthermore, we will put in our INPUT statement the following symbols @ @. In SAS this set of symbols allows you to put the data for more than one subject on the same line.

In SAS, as with the other packages, there are certain rules for variable names. Each variable name must begin with a letter and be 8 or less characters. The variable name can contain numbers, but *not* special characters or an imbedded blank(s). For example, I.Q., $x1 + x2$, and SOC CLAS, are not valid variable names. We have special characters in the first two names (periods in I.Q. and the + in $x1 + x2$) and there is an embedded blank in the abbreviation for social class.

Our INPUT statement is as follows:

```
INPUT LOCUS ACHMOT ACHLANG @@;
```

Following the INPUT statement there is a LINES statement, which tells SAS that the data is to follow. Thus, the first three statements here setting up the data look like this:

```
DATA LOCUS;
```

```
INPUT LOCUS ACHMOT ACHLANG @@;
LINES;
```

Recall that the next structural part of a SAS program is the set of data lines. Remember there are dime variables, so we have three scores for each subject. We will put the scores for three subjects on each data line. Adding the data lines to the above three statements, we now have the following part of the SAS program:

```
DATA LOCUS;
INPUT LOCUS ACHMOT ACHLANG @@;
LINES;
11 23 31 13 25 38 21 28 29
21 34 28 14 36 37 29 20 37
17 24 39 19 30 39 23 28 41
```

The first 3 scores (11, 23, and 31) are the scores on locus of control, achievement motivation, and achievement in language for the first subject; the next 3 numbers (13, 25, and 38) are the scores on these variables for subject 2; etc.

Now we come to the last structural part of a SAS program, calling up some SAS procedure(s) to do whatever statistical analysis(es) we desire. In this case we want correlations, and the SAS procedure for that is called CORR. Also, as mentioned earlier, we should always print the data. For this we use PROC PRINT. Adding these lines we get our complete SAS program:

```
DATA LOCUS;
INPUT LOCUS ACHMOT ACHLANG @@;
LINES;
11 23 31 13 25 38 21 28 29
21 34 28 14 36 37 29 20 37
17 24 39 19 30 39 23 28 41
PROC CORR;
PROC PRINT;
```

Note that there is a semicolon at the end of each statement, but *not* for the data lines.

In Table 1.1 we present some of the basic rules of the control language for SAS, and in Table 1.2 give the complete SAS control lines for obtaining descriptive statistics, for obtaining a set of correlations (this is the example we just went over in detail), and for obtaining both the independent and dependent samples *t* tests. Although the rules are basic, they are important. For example, failing to end a statement in SAS with a semicolon or using a variable name longer than 8 characters will cause the program to terminate. The four sets of control lines in Table 1.2 show the structural similarity of the control line flow for different types of analyses. Notice in each case we start with the DATA statement, then an INPUT state-

TABLE 1.1
Some Basic Elements of the SAS Control Language

Non-columned oriented. Columns only become relevant when using column input for the data.

SAS statements give instructions. Each statement *must* end with a semicolon.

Structurally a SAS program is composed of three fundamental blocks: (1) statements setting up the data, (2) the data lines and (3) procedure (PROC) statements—procedures are SAS computer programs which read the data and do various statistical analyses.

DATA SETUP

First there is the DATA statement, where you are creating a data set. The name for the data set must begin with a letter and be 8 or less characters.

Then there is the INPUT statement, where the variables are named and the format of the data is specified.

Variable names must be 8 or less characters, must begin with a letter, and cannot contain special characters.

We can use *column* input, where we indicate what column(s) the score for a variable is. If the variable is non-numeric then we need to put a \$ after the variable name.

Example

Suppose we have a group of subjects measured on IQ, attitude toward education and grade point average (GPA), and will label them as M for male and F for female. SEX \$ 1 IQ 3–5 ATTITUDE 7–8 GPA 10–12.2

This tells SAS that sex (M or F) is in column 1, IQ is in columns 3 through 5, attitude in columns 7 to 8, and grade point average in columns 10 to 12. The .2 is to insert a decimal point *before* the last two digits.

If we are using free format then the scores for the variables do *not* have to be in specific columns, they simply need to be separated from each other by at least one blank.

The LINES statement follows the DATA and INPUT statements and precedes the data lines

More than one statement can go on the same line, although for readability we recommend putting statements on separate lines. If we wish to do analysis on only some of the variables in the INPUT statement, then this is indicated in a VAR (abbreviation for variable) statement. For example, if we had 6 variables on the INPUT statement (X1 X2 X3 X4 X5 X6) and only wished to compute correlations for the first 3, then we would insert VAR X1 X2 X3; after the PROC CORR statement.

Statistics for subgroups of subjects are obtained with the BY statement. Suppose we want the correlations for males and females on variables X, Y and Z. If the subjects have not been sorted on sex, then we sort them first using PROC SORT, and the control lines are

```
PROC CORR;  
PROC SORT;  
BY SEX;
```

TABLE 1.2
SAS Control Lines for Obtaining Set of Correlations, Descriptive Statistics,
and Independent and Dependent T Tests

<u>CORRELATIONS</u>	<u>T TEST</u>
① DATA LOCUS; ② INPUT LOCUS ACHMOT ACHLANG @@; ③ LINES; 11 23 31 13 25 38 21 28 29 21 34 28 14 36 37 29 20 27 17 24 39 19 30 39 23 28 41 PROC CORR; ④ PROC PRINT;	DATA ATTITUDE; ⑤ INPUT TREAT \$ ATT @@; LINES; C 82 C 95 C 89 99 C 87 C 79 C 98 C 86 T 94 T 97 T 98 T 93 T 96 T 99 T 88 T 92 T 94 T 95 T 92 T 97 T 96 T 90 T 89 PROC TTEST; ⑥ CLASS TREAT; PROC PRINT;
<u>MEANS AND STANDARD DEVIATIONS</u>	<u>DEPENDENT SAMPLES T TEST</u>
DATA MEANS; INPUT DRINK \$ TACTUAL @@; LINES; A 34 A 26 A 18 A 26 A 9 A 28 A 14 A 33 A 43 A 50 NA 15 NA 2 NA 23 NA 7 NA 18 NA 13 NA 9 NA 23 NA 8 NA 16 PROC MEANS; ⑦ BY DRINK;	DATA COFFEE; INPUT PRODWCB PRODWITH @@; ⑧ DIFF = PRODWITH-PRODWCB; LINES; 23 28 35 38 29 29 33 37 43 42 32 30 ⑨ PROC MEANS N MEAN T PRT; VAR DIFF;

① Here we are giving a name to the data set. Remember it must be eight or less letters and must begin with a letter. Note that there is a semicolon at the end of the line, and at the end of *every* line for all 4 examples (except for the data lines).

② Note that the names for the variables all begin with a letter and are less than or equal to 8 characters. The double @@ is needed in order to put the data for more than one subject on the same data line; here we have data for 3 subjects on each line.

③ When the data is part of the control lines, as here, then this LINES command always precedes the data.

④ PROC (short for procedure) CORR yields the correlations, and PROC PRINT gives a listing of the data.

⑤ The \$ after TREAT is used to denote a non-numeric variable; note in the data lines that TREAT is either C(control) or T(treatment).

⑥ We call up the t test procedure and tell it that TREAT is the grouping variable.

⑦ This BY statement yields means and standard deviations for each of the subgroups defined by DRINK (alcoholics & non-alcoholics)

⑧ We create the difference variable (DIFF) on which the analysis is done.

⑨ Procedure MEANS is used, with MEAN, T, and PRT yielding the mean on the difference variable, *t* is test statistic and PRT is the tail probability.

ment (naming the variables being read in and describing the format of the data), and then the LINES statement preceding the data. Then, after the data, one or more PROC statements are used to perform the wanted statistical analysis, or to print the data (PROC PRINT).

These 4 sets of control lines serve as useful models for running analyses of the same type, where only the variable names change and/or the names and number of variables change. For example, suppose you want all correlations on 5 attitudinal variables (call them x_1 , x_2 , x_3 , x_4 , and x_5). Then the control lines are:

```
DATA ATTITUDE;
INPUT X1 X2 X3 X4 X5 @@;
LINES;
```

```
DATA LINES
```

```
PROC CORR;
PROC PRINT;
```

where the data lines have just been indicated schematically.

In Table 1.3 we present some of the basic elements of the SPSS control language, and in Table 1.4 give the complete SPSS control lines for descriptive statistics, correlations, and the t tests for independent and dependent samples. Some of the common errors committed in running SPSS programs are (1) using invalid variable names, (2) failing to indent for a subcommand, and (3) not starting a command in column 1.

It should be understood that although we give some important basic elements of the packages in Tables 1.1 and 1.3, and present *complete* control lines for various types of analyses in this text, our treatment is in no sense a substitute for the SAS and SPSS manuals. All the contingencies one might encounter in a practical problem can't be covered in this text. One final important point before we leave the packages: The examples in this book were run on SPSS for Windows 12.0 or SAS. It is possible if you are running a different release of SPSS or SAS that things may change a bit. Your instructor can help you with this.

1.8 SPSS FOR WINDOWS—RELEASE 12.0

A fantastic bargain, in my opinion, is the SPSS GRADUATE PACK FOR WINDOWS 12.0, which comes on a compact disk and sells at a university for students for about \$190. It is important to understand that you are getting the full package, not a student version. For this release, as they note, Windows 98/2000 Professional or NT 4.0 Workstation, ME and XP are required, along with 128 MB

TABLE 1.3
Some Basic Elements of the SPSS Control Language

SPSS operates on commands and subcommands

It is column oriented to the extent that each command begins in column 1 and continues for as many lines as needed. All continuation lines are indented at least one column.

Examples of Commands: TITLE, DATA LIST, BEGIN DATA

The title can be put in apostrophes, and can be up to 60 characters.

All subcommands begin with a keyword followed by an equals sign, then the specifications, and are terminated by a slash.

Each subcommand is indented at least one column.

The subcommands are further specifications for the commands.

For example, if the command is DATA LIST, then

DATA LIST FREE involves the subcommand FREE which indicates the data will be in free format.

Names for variables must be eight or less characters.

They must begin with a letter, or one of the following characters: @, # or \$.

FREE format—the variables must be in the same order for each case but do not have to be in the same location. Also, multiple cases can go on the same line, with the values for the variables separated by blanks or commas.

When the data is part of the command file, then the BEGIN DATA command precedes the data and the END DATA follows the last line of data.

The LIST command can be used to list the data.

We can use the keyword TO in specifying a set of consecutive variables, rather than listing all the variables. For example, if we had the six variables X1,X2,X3,X4,X5,X6, the following subcommand are equivalent:

VARIABLES = X1,X2,X3,X4,X5,X6/ or VARIABLES = X1 TO X6/

of RAM and 200 MB of hard disk space. If you have purchased a computer in the last 3 years these requirements should not be a problem. Statistical analysis is done on data, so getting data into SPSS or SAS is crucial. We discuss this next.

1.9 DATA FILES

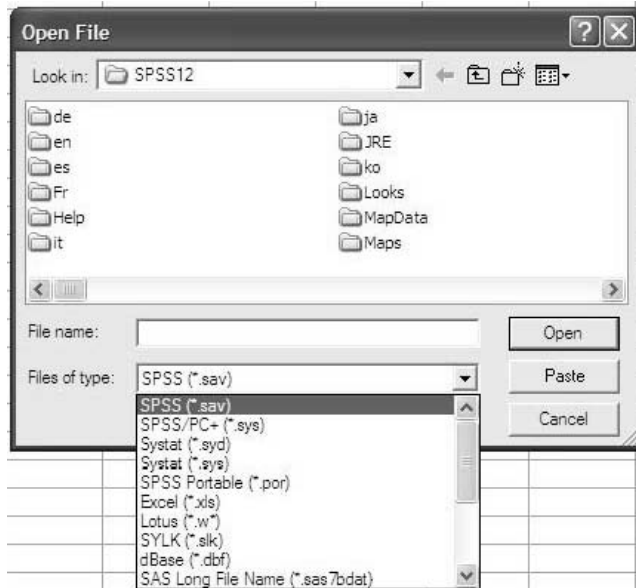
As noted in the SPSS BASE12.0 USER'S GUIDE (2003, p.19), "Data files come in a variety of formats, and this software is designed to handle many of them, including:

TABLE 1.4
SPSS Control Lines for Obtaining a Set of Correlations Descriptive
Statistics, and Independent and Dependent T Tests

<u>CORRELATIONS</u>	<u>T TEST</u>
TITLE 'CORRELATIONS FOR 3 VARIABLES'.	TITLE 'T TEST'.
① DATA LIST FREE/LOCUS ACHMOT ACHLANG.	DATA LIST FREE/TREAT ATT.
② BEGIN DATA.	BEGIN DATA.
11 23 31 13 25 38 21 28 29	⑥ 1 82 1 95 1 89 1 99
11 34 28 14 36 37 29 20 37	1 87 1 79 1 98 1 86
17 24 39 19 30 39 23 28 41	2 94 2 97 2 98 2 93
END DATA.	2 96 2 99 2 88 2 92
③ CORRELATIONS VARIABLES = LOCUS ACHMOT	2 94 2 95 2 92 2 97
ACHLANG/	2 96 2 90 2 89
PRINT = TWOTAIL/	END DATA.
④ STATISTICS = DESCRIPTIVES/.	⑤ LIST.
	⑦ T-TEST GROUPS = TREAT (1,2) / VARIABLES = ATT/.
<u>MEANS AND STANDARD DEVIATIONS</u>	<u>DEPENDENT SAMPLES T TEST</u>
TITLE 'DESCRIPTIVE STATISTICS'.	TITLE 'COFFEE BREAK'
DATA LIST FREE/DRINK TACTUAL.	DATA LIST FREE/PWO PWITH.
VALUE LABELS DRINK 1 'ALCOHOLIC'	BEGIN DATA.
2 'NON ALCOHOLIC'.	23 28 35 38 29 29
BEGIN DATA.	33 37 43 42 32 30
1 34 1 26 1 18 1 26 1 9	END DATA.
1 28 1 14 1 33 1 43 1 50	⑨ T-TEST PAIRS = PWO PWITH/.
2 15 2 2 2 23 2 7 2 18	
2 13 2 9 2 23 2 8 2 16	
END DATA.	
⑧ MEANS TABLES = TACTUAL BY DRINK/.	
① The FREE on this DATA LIST command is a further specification, indicating that the data will be in free format.	
② When the data is part of the command file, it is preceded by BEGIN DATA and terminated by END DATA.	
③ This VARIABLES subcommand specifies the variables to be analyzed.	
④ This yields the means and standard deviations for all variables.	
⑤ This LIST command gives a listing of the data.	
⑥ The first number for each pair is the group identification and the second is the score for the dependent variable. Thus, 82 is the score for the first subject in group 1 and 97 is the score for the second subject in group 2.	
⑦ The t test procedure is called and the number of levels for the grouping variables is put in parentheses.	
⑧ The MEANS procedure calculates means and variances for a dependent variable(s) over subgroups defined by one or more classification variables. The TABLES subcommand is used to indicate for which variables the means and variances are desired.	
⑨ The PAIRS subcommand names the variables being compared.	

- Spreadsheets created with Lotus1–2–3 and Excel
- Database files created with dBASE and various SQL formats
- Tab delimited and other types of ASCII text files
- Data files in SPSS format created on other operating systems
- SYSTAT data files
- SAS data files

As the screen below shows, one can easily import files of different types into SPSS for analysis:



We illustrate for an EXCEL and an SPSS file. As the above screen indicates, one needs to tell the software *where* the file is located and what *type* of file it is. If it is an EXCEL file (stored in MY DOCUMENTS), then one would select MY DOCUMENTS and EXCEL for type of file. If it is an SPSS file, stored in SPSS12, then select that location and SPSS(*SAV) for type of file. We illustrate for an SPSS file, which we saved as INTERMCH1. The file is:

SPSS FILE

23

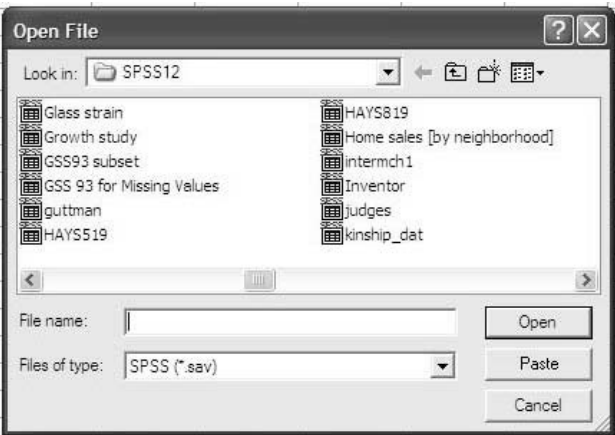
25

27

29

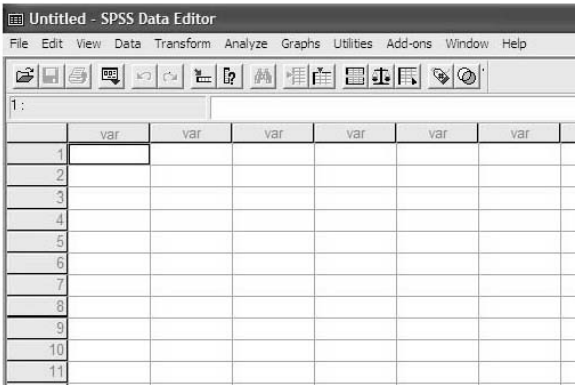
31

For the SPSS file the screen would be



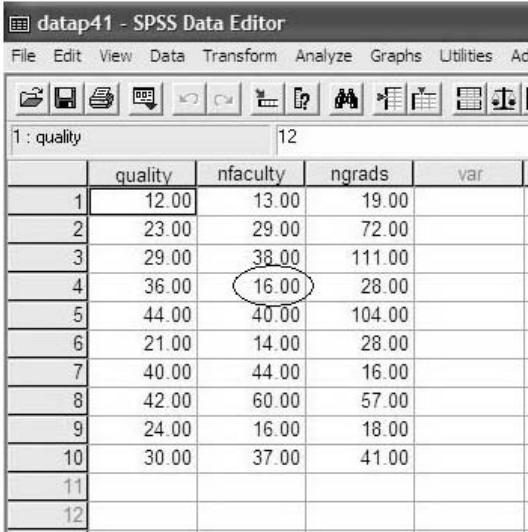
1.10 DATA ENTRY

When SPSS is opened the data editor provides a spreadsheet like editor for creating and editing data files. In this section we illustrate creating a data set within SPSS. The data set we create has 3 variables and 10 cases. In the editor cases are rows and variables are columns. The data editor is shown below:



The first number we wish to enter is 12. Press the forward arrow key and you will move laterally to the next column. There you enter the value for the 2nd variable, i.e., 13. Press the forward arrow key again and enter 19. Now, you press TAB and the box will go *automatically* to the first position in the second row. Punch in 23 and press the TAB key. Now, punch in 29 and press the TAB key again. Then en-

ter 72 and press the TAB key again. The box will go automatically to the first position in the third row. When you are done punching in all 10 cases, the screen looks as follows:



The screenshot shows the SPSS Data Editor window titled 'datap41 - SPSS Data Editor'. The menu bar includes File, Edit, View, Data, Transform, Analyze, Graphs, Utilities, and Add. The toolbar contains icons for file operations, editing, and analysis. The data grid shows 12 rows and 4 columns. The first column is labeled '1 : quality' and the second column is labeled '12'. The data is as follows:

	quality	nfaculty	ngrads	var
1	12.00	13.00	19.00	
2	23.00	29.00	72.00	
3	29.00	38.00	111.00	
4	36.00	16.00	28.00	
5	44.00	40.00	104.00	
6	21.00	14.00	28.00	
7	40.00	44.00	16.00	
8	42.00	60.00	57.00	
9	24.00	16.00	18.00	
10	30.00	37.00	41.00	
11				
12				

We have skipped a step here. Originally, the program assigns generic names to the variables. By switching to the VARIABLE VIEW we have given the above names to the variables. Switching back to the DATA VIEW we obtain the above result.

1.11 EDITING A DATASET

Changing a Cell Value

Suppose we wished to change the circled value to 23. Move to that cell. Enter the 23 and press ENTER. The new value appears in the cell. It is as simple as that.

Inserting a Case

Suppose we wished to insert a case after the 7th subject. How would we do it? As the guide points out:

1. Select any cell in the case (row) *below* the position where you want to insert the new case.
2. From the menus choose:

DATA INSERT CASE

A new row is inserted for the case and all variables receive the system-missing value. It would look as follows:

Insert art from p. 40 of previous edition

18 x 20 picas

Suppose the new case we typed in was 35 17 63.

Inserting a Variable

Now we wish to add a variable after NFACULTY. How would we do it?

1. Select any cell in the variable (column) to the *right* of the position where you want to insert the new variable.
2. From the menus choose:

DATA
INSERT A VARIABLE

When this is done, the data file in the editor looks as follows:

Insert art from p. 41 of previous edition

21 x 20 picas

Deleting a Case

To delete a case is also simple. Click on the row (case) you wish to delete. The entire row is highlighted. From the menus choose:

EDIT

CLEAR

The selected row (case) is deleted and the cases below it move it up. To illustrate, suppose for the above data set we wished to delete case 4 (row 4). Click on 4 and choose EDIT and CLEAR. The case is deleted, and we are back to 10 cases, as shown below:

Insert art from p. 42 of previous edition

22.5 x 20 picas

1.12 SPLITTING AND MERGING FILES

Split file analysis splits the data file into separate groups for analysis, based on the values of the grouping variable (there can be more than one). We will find this useful in Chapter 2 on assumptions when we wish to obtain the z scores *within* each group. To obtain a split file analysis, click on DATA and then on SPLIT FILE from the dropdown menu. Select the variable on which you wish to divide into groups and then select ORGANIZE OUTPUT BY GROUPS.

Merging data files can be done in two different ways: (1) merging files with the same variables and different cases, and (2) merging files with the same cases but different variables. SPSS gives the following marketing example for the first case. For example, you might record the same information for customers in two different sales regions and maintain the data for each region in separate files. We will give an example to illustrate how one would merge files with the same variable and different cases. As the guide notes, open one of the data files. Then, from the menus choose:

DATA
MERGE FILES
ADD CASES

Then select the data file to merge with the open data file.

Example

To illustrate the process of merging files, we consider two small, artificial data sets. We denote these data sets by MERGE1 and MERGE2, respectively, and they are shown below:

Insert art from p. 43 of previous edition

24 x 24 picas

As indicated above, we open MERGE1 and then select DATA and MERGE FILES and ADD CASES from the dropdown menus. When we open MERGE2 the ADD CASES window appears:

Insert art from p. 44 of previous edition (top half)

29.5 x 20 picase

When you click on the OK the merged file appears, as given below:

Insert art from p. 44 of previous edition (bottom half)

21 x 18 picas

1.13 TWO WAYS OF RUNNING ANALYSES ON SPSS

Point and Click

Bring the data into the editor. Click on ANALYZE and scroll down to analysis desired.

Syntax Editor

- Click on FILE, NEW AND SYNTAX. A blank screen will appear.
- Type in the syntax.
- To run, click on RUN and then scroll down to ALL.

To illustrate both methods of doing an analysis we use the t test data from Table 1.4.

For the point and click method we would first bring the data into the spreadsheet editor. Then we click on ANALYZE, scroll down to COMPARE MEANS, and across to INDEPENDENT SAMPLES T TEST.

To use the syntax editor for analysis, we first need to get to the syntax editor. This we do with FILE—NEW—SYNTAX. A blank sheet appears.

We simply type in the syntax; then click on RUN and then ALL.

Both methods will, of course, yield the same results.

1.14 SPSS OUTPUT NAVIGATOR

The Output Navigator was introduced in SPSS for Windows (7.0) in 1996. It is very nice. A survey researcher is conducting a pilot study on a 12 item scale to check out possible ambiguous wording, whether any items are sensitive, whether they discriminate, etc. She administers the scale to 16 subjects. The items are scaled from 1 to 5, with 1 representing strongly agree and 5 representing strongly disagree. The first 8 subjects are male and the last 8 are female. There is some missing data, which is coded as a 0. She wishes to compare males and females on 3 subtests (SUBTEST1, SUBTEST2, SUBTEST3), and also to determine the internal consistency of these subtests. We will illustrate only *some* of the things that can be done with output for the above survey example. First, the entire command syntax for running the analysis is presented below:

```
TITLE 'SURVEY RESEARCH WITH MISSING DATA'.
DATA LIST FREE/ID I1 I2 I3 I4 I5 I6 I7 I8 I9 I10 I11 I12 SEX.
BEGIN DATA.
```

```

1 1 2 2 3 3 1 1 2 2 1 2 2 1
2 1 2 2 3 3 3 1 2 2 1 1 1 1
3 1 2 1 3 3 2 3 3 2 1 2 3 1
4 2 2 4 2 3 3 2 2 3 3 2 3 1
5 2 3 2 4 2 1 2 3 0 3 4 0 1
6 2 3 2 3 3 2 3 4 3 2 4 2 1
7 3 4 4 3 5 2 2 1 2 3 3 4 1
8 3 2 3 4 4 3 4 3 3 3 4 2 1
9 3 3 4 2 4 3 3 4 5 3 5 3 2
10 4 4 5 5 3 3 5 4 4 4 5 3 2
11 4 4 0 5 5 5 4 3 0 5 4 4 2
12 4 4 4 5 5 4 3 3 5 4 4 5 2
13 4 4 0 4 3 2 5 1 3 3 0 4 2
14 5 5 3 4 4 4 4 5 3 5 5 3 2
15 5 5 4 5 3 5 5 4 4 5 3 5 2
16 5 4 3 4 3 5 4 4 3 2 2 3 2

```

END DATA.

LIST.

MISSING VALUES ALL (0).

COMPUTE SUBTEST1 = I1+I2+I3+I4+I5.

COMPUTE SUBTEST2 = I6+I7+I8+I9.

COMPUTE SUBTEST3 = I10+I11+I12.

RELIABILITY VARIABLES = I1 TO I12/

SCALE(SUBTEST1) = I1 TO I5/

SCALE(SUBTEST2) = I6 TO I9/

SCALE(SUBTEST3) = I10 I11 I12/

STATISTICS = CORR/.

T-TEST GROUPS = SEX(1,2)/

VARIABLES = SUBTEST1 SUBTEST2 SUBTEST3/.

This is run from the command syntax window by clicking on RUN and then on ALL. The first thing you want to do is save the output. To do that click on FILE and then click on SAVE AS from the dropdown menu. Type in a name for the output (we will use MISSING), and then click on OK. The output, in the navigator, appears as follows:

Insert art from p. 46 of previous edition

31 x 24 picas

As shown above, the output is divided into two panes. The left pane gives in outline form the analysis(es) that have been run, and the right pane has the statistical contents. To print the entire output, simply click on FILE and then click on PRINT from the dropdown menu. Select how many copies you want and click on OK. It is also possible to print only part of the output. I will illustrate. Suppose we wished to print only the reliability part of the output. Click on that in the left part of the pane; it is highlighted (as shown in the figure below). Click on FILE and PRINT from the dropdown menu. Now, when the print window appears click on SELECTION and then OK. Only the reliability part of the output will be printed.

Insert art from p. 47 of previous edition (top half)

21 x 23 picas

Insert art from p. 47 of previous edition (bottom half)

28 x 16 picas

Insert art from p. 48 of previous edition

It is also easy to move and delete output in the output navigator. Suppose for the missing data example we wished to move the corresponding to LIST to just above the t Test. We simply click on the LIST in the outline pane and drag it (holding the mouse down) to just above the t test and then release.

To delete output is also easy. Suppose we wish to delete the LIST output. Click on LIST. To delete the output one can either hit DEL (delete) key on the keyboard, or click on EDIT and then click on DELETE from the dropdown menu.

As mentioned at the beginning of this section, there are many, many other things one can do with output.

1.15 SAS AND SPSS OUTPUT FOR CORRELATIONS,
 DESCRIPTIVES, AND *t* TESTS

In Table 1.5 we present SPSS Windows 12.0 printout for the correlations and SAS printout for the descriptives statistics. Table 1.6 presents SPSS for Windows 12.0 screens for the *t* test for independent samples. Table 1.7 has the SPSS for Windows 12.0 printout for the independent samples *t* test and the SAS printout for the dependent samples *t* test.

TABLE 1.5
Correlations and Descriptive Statistics From SPSS
for Windows and SAS

Insert art from p. 49 of previous edition

27 x 25 picas

TABLE 1.6
SPSS Windows Screens for Running *t* Test for Independent Samples

Insert art from p. 50 of previous edition

30 x 19.5

Insert art from p. 50 of previous edition

?? X 14 picas

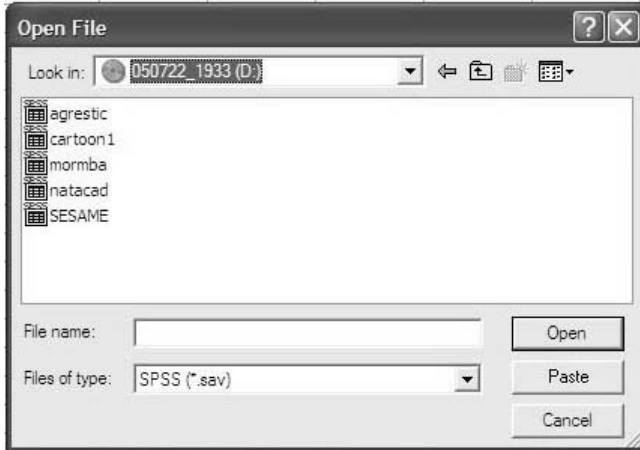
TABLE 1.7
t Test for Independent Samples From SPSS Windows and *t* Test
for Correlated Samples From SAS

Insert art from p. 51 of previous edition

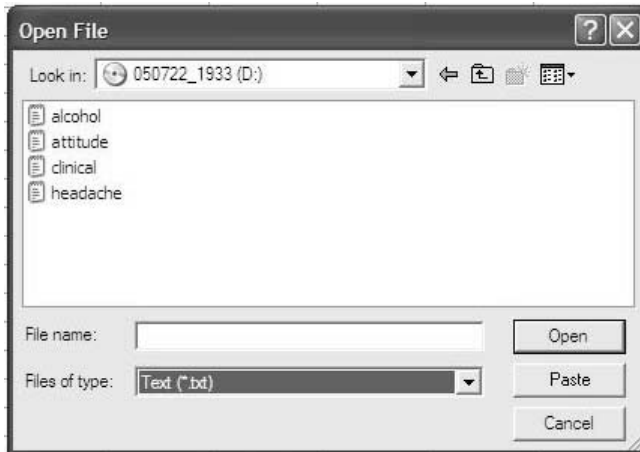
X 38.5 picas depth

1.16 DATA SETS ON COMPACT DISK

There are 5 SPSS data files on the compact disk, and 4 ASCII (text) data files on the disk. To access the SPSS files change LOOK IN to the compact disk icon and FILE TYPE to SPSS(*.sav). When this is done, the screen will appear as follows:



To access the ASCII(text) files leave LOOK IN as the compact disk icon, but change FILE TYPE to TEXT. When this is done, the screen will look as follows:



When you double click on a SPSS file, the file will go right into the spreadsheet editor ready for analysis. For the ASCII files things are a bit more complicated. When one double clicks on an ASCII file, the TEXT WIZARD will appear. This is documented in SPSS BASE 12.0 USER'S GUIDE (2003, PP37–47 and SPSS BASE 13.0 USER'S GUIDE (2004, PP 39–49). This procedure can read a variety of files. For our purpose just press NEXT several times. In the final step (step 6) press FINNISH, and the data file will appear in the spreadsheet editor ready for analysis.

EXERCISES

1. An advertisement in the paper claims that the average pay at Smith Industries, a small factory, is \$22,000. You are currently making \$15,000 and decide to apply for a job there. Subsequently you find out that most people at Smith also make \$15,000, and you are upset about the ad. You later determine that the salary structure at Smith Industries is as follows:

50 workers	\$15,000 each
Managers of the two divisions at Smith	\$35,000 each
Two executives	\$70,000 each
Owner	\$250,000

- (a) Why was the \$22,000 figure in the paper so misleading?
- (b) Which measure(s) of central tendency should have been used to convey a more accurate picture of the salaries at Smith?
2. Suppose Mr. Jones had administered the same math test to each of his two eighth grade classes, with the results shown below:

	<i>Class 1</i>	<i>Class 2</i>
Size	20	40
Mean	60	80

Mr. Jones then naively computed the average for all students by taking the average of the above two means, yielding 70.

- (a) Intuitively, why is this not correct?
- (b) The correct formula for finding the combined mean is to use a weighted average:

$$\bar{x}_c = (n_1\bar{x}_1 + n_2\bar{x}_2)/(n_1 + n_2)$$

where n_1 and n_2 are the respective group sizes and \bar{x}_1 and \bar{x}_2 are the group means. Plugging the above numbers into this formula yields the correct overall mean of 73.33.

Now, let x_1, x_2, \dots, x_{n1} represent the subjects scores in group 1 and let x_1, x_2, \dots, x_{n2} represent the subjects scores in group 2. Prove that the formula for the combined mean is as given above.

HINT: Start with the definition for the mean for all subjects combined:

$$\bar{x}_c = \frac{(x_1 + x_2 + \dots + x_{n1}) + (x_1 + x_2 + \dots + x_{n2})}{n_1 + n_2}$$

3. An investigator runs a t test for independent samples on two groups of subjects (45 subjects in group 1 and 35 in group 2). She notes that the distributions of scores are quite positively skewed in both groups. Should she be concerned about this?
4. (a) Suppose that in a hospital each patient's pulse is taken in the morning, at noon, and in the evening. For two patients, on a given day, the average pulse readings are both 74. The records for that day show the following:

	<i>Morning</i>	<i>Noon</i>	<i>Evening</i>	<i>Mean</i>
Patient A	72	76	74	74
Patient B	72	91	59	74

Are the clinical implications for these two patients the same? Explain.

(b) You are a scout for the Boston Celtics professional basketball team and are looking for a guard. From scouting reports you focus in on two guards from Duke and UCLA, schools that play on a similar level of competition. It is noted that each guard averaged 20 points over all games in his senior year, so you decide to examine their performance game by game:

UCLA guard: 21, 18, 19, 23, 25, 20, 22, 17, 23, 24 etc.

Duke guard: 13, 35, 28, 11, 8, 40, 22, 31, 15, 29 etc.

- (a) Which guard might you prefer, and why?
 - (b) What is the main point that each of the two parts of this exercise illustrates?
5. (a) Suppose that $c = 3$, $x_1 = 5$, $x_2 = 8$, $x_3 = 1$, and $x_4 = 7$. Evaluate the following:

$$\sum_{i=1}^4 cx_i$$

(b) Prove that the variance of a constant times a variable is equal to c^2 times the variance of x ; that is, prove that

$$s_{cx}^2 = c^2 s_x^2$$

Hint: The scores on cx may be represented as: cx_1, cx_2, \dots, cx_n . Apply the definitional formula for variance to this set of scores and mathematically rearrange.

(c) Suppose there are 10 subjects in each of three groups. The means for the groups are $\bar{x}_1 = 4.1$, $\bar{x}_2 = 6.3$ and $\bar{x}_3 = 8.5$. Evaluate the following:

$$\sum_{i=1}^3 10(\bar{x}_i - \bar{x})^2, \text{ where } \bar{x} \text{ is the grand mean for all the subjects.}$$

6. A team of researchers is comparing two diets, a behavior modification approach and the Beverly Hills diet, in their effect on weight loss for a group of overweight women. Suppose that the data for the 10 subjects in each diet is as follows:

Behavioral Modification		Beverly Hills	
10	14	8	12
15	17	16	10
11	32	13	7
22	15	4	15
8	9	10	1

Label the independent variable here DIET and the dependent variable WGTLOSS. Use column format, with group identification (1 or 2) in column 1 and weight loss in columns 3 and 4. Show the complete SAS control lines for running the t test for independent samples on this data.

7. A researcher has the following heights and weights on 14 men:

	Subject													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Height	67	66	63	61	68	69	69	70	71	73	75	74	77	71
Weight	148	161	152	145	169	162	170	183	174	115	205	186	233	158

The heights are given in inches.

- (a) Compute the correlation between height and weight. What conclusion would you draw?
- (b) Check the data to see if there is an outlier.
- (c) Compute the correlation without the outlier. Now, what do you conclude?

8. A worker in a neighborhood clinic wishes to assess the impact of showing an educational film on patient compliance in taking an antihypertension medication. The diastolic blood pressure is the dependent variable here. During the study the medication dosage is kept constant. Blood pressure was measured one week prior to the film, then the film was shown, and the blood pressure was measured again three weeks later. The data are:

	<i>Patient</i>									
	1	2	3	4	5	6	7	8	9	10
<i>Before</i>	110	105	98	100	89	82	113	102	101	118
<i>After</i>	100	95	88	92	83	86	100	101	96	112

- (a) Denote the diastolic blood pressure by *DIASTOL* and the treatment (independent variable) by *FILM*. Show the complete set of control lines for running the *t* test for dependent samples on SPSS to determine whether there was a significant change in the blood pressure.
9. Run a *t* test for independent samples with *TREAT* as the grouping variable and attitude (*ATT*) as the dependent variable. Use either SAS (Table 1.2) or SPSS (Table 1.4).
10. Run a *t* test for dependent samples with SAS (Table 1.2) or SPSS (Table 1.4).
11. Consider the following small data set:

	<i>X</i>	<i>Y</i>	<i>Z</i>
1	12	13	15
2	11	9	8
3	7	5	3
4	2	4	6
5	1	5	3
6	8	9	7

Edit this data set using SPSS or SAS in the following ways:

- (a) Change 13 to 25.
 (b) Insert a case after Case 3
 (c) Insert a variable after *Y*.

APPENDIX

OBTAINING THE MEAN AND VARIANCE

ON THE TI-30Xa CALCULATOR

We consider the data for the children in group 1 (example in Section 1.2), which is as follows:

10, 13, 7, 12, 13, 11, 8, 14, 9, 12

TI-30Xa

STEP	DISPLAY
1. Enter 10 and press $\Sigma+$	n=1 (indicates 1 data point has been entered)
2. Enter 13 and press $\Sigma+$	n=2 (indicates 2 data points entered)
3. Enter 7 and press $\Sigma+$	n=3
4. Enter the remaining 7 data points, and at that point the display will show n=10, indicating that 10 data points have been entered.	
5. Press 2ND and then X^2	10.9 (this is the mean)
6. Press 2ND and then \sqrt{X}	2.33095 (standard deviation)
7. Press X^2	5.4333 (variance)

One Way Analysis of Variance

CONTENTS

- 2.1 Introduction
- 2.2 Rationale for ANOVA
- 2.3 Numerical Example
- 2.4 Expected Mean Squares
- 2.5 MS_W and MS_B as Variances
- 2.6 A Linear Model for the Data
- 2.7 Assumptions in ANOVA
- 2.8 The Independence Assumption
- 2.9 ANOVA on SPSS and SAS
- 2.10 Post Hoc Procedures
- 2.11 Tukey Procedure
- 2.12 The Scheffé Procedure
- 2.13 Heterogeneous Variances and Unequal Group Sizes
- 2.14 Measures of Association (Variance Accounted For)
- 2.15 Planned Comparisons
- 2.16 Test Statistic for Planned Comparisons
- 2.17 Planned Comparisons on SPSS and SAS
- 2.18 The Effect of an Outlier on an ANOVA
- 2.19 Multivariate Analysis of Variance
- 2.20 Summary
- Appendix

2.1 INTRODUCTION

One of the statistical tests encountered in introductory statistics courses is the t test for independent samples. This test is appropriate when comparing two groups of subjects on a single dependent variable. Three classical applications of this test are given below:

1. Comparing a treatment group against a control group.
2. Comparing the relative efficacy of treatment 1 vs. treatment 2
3. Comparing two intact groups (such as males and females or two social classes) on some dependent variable.

However, in many situations we may wish to compare more than two groups simultaneously on a dependent variable. In these cases a different statistical technique, called analysis of variance (ANOVA), is needed. We consider 7 examples below:

1. A counselor wishes to compare the effectiveness of two types of counseling* (Rogerian and Adlerian) on changing the attitude of low achieving high school students toward school. She also has a control group in her study. Thus, there are 3 groups being compared on the dependent variable of attitude toward school.
2. A psychologist wants to determine if five drugs have a differential effect on reaction time (dependent variable) for 100 subjects, 20 of which have been randomly assigned to each drug.
3. A dietician wishes to discover whether four diets produce differential weight loss (the dependent variable) for 80 overweight women. Here diets are the treatments, and we have four groups being compared.
4. A researcher for a school district wishes to determine whether the reading achievement on a standardized test differs on the average for six elementary schools in similar socioeconomic areas. Here we have six groups (the schools) being compared on reading achievement (the dependent variable).
5. A marketing researcher wishes to determine if shelf location has an effect on volume of sales of a product. If there are 4 shelf locations (from high to low), then we have a 4 group ANOVA, with sales as the dependent variable.

*This chapter deals with what is called *fixed effects* ANOVA, since counseling methods, teaching methods, diets, etc. are generally not randomly sampled from some population of methods or diets. Thus, our inferences are fixed to the counseling methods, teaching methods or diets under consideration. Further elaboration of this point is given in Section 4.7, where we distinguish between fixed and random effects ANOVA.

6. We wish to determine whether violent crime varies in different regions of the country. If there are 7 different regions being compared, then we have a 7 group ANOVA.
7. Doughnuts absorb fat in various amounts when they are cooked. Suppose an experiment is set up involving three types of fat: peanut oil, corn oil, and lard. We would have a 3 group ANOVA, with amount of fat absorbed as the dependent variable.

Let us review some of the basic terminology concerned with hypothesis testing before considering how to do an ANOVA. Recall that with the t test we talked about testing a null hypothesis versus some alternative hypothesis, which looked like this:

$$H_0: \mu_1 = \mu_2 \text{ (population means are equal)}$$

$$H_1: \mu_1 \neq \mu_2$$

Why is it called the *null* hypothesis? Because to say that the population means are equal is equivalent to saying that their difference is null, i.e., that $\mu_1 - \mu_2 = 0$. Also, in testing the null hypothesis the notion of testing at some level of significance was encountered. What does it mean to do a t test at the $\alpha = .05$ level of significance? This means we are taking a 5% chance of rejecting the null hypothesis when it is true, that is, saying the groups differ when in fact they do not. Level of significance is also called the probability of making a type I error.

Notice that the alternative hypothesis (H_1) for the t test is very simple, since either the two groups are equal or they differ. In analysis of variance (for k groups) the alternative is much more complex. What is the null hypothesis for a one way analysis of variance with k groups? It is that the k population means are equal:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_k$$

The alternative hypothesis here is more complicated than that for the t test. Let us consider the four group case to illustrate. If we reject the null hypothesis it could be for various reasons. It might be because only two groups are different, or it might be because only 3 groups are different, or because group 1 differs only from groups 3 and 4, or because all 4 groups are different, etc. How can we characterize all these possibilities into an alternative hypothesis? Notice that in all the above cases *at least two* of the groups differed. Thus, a way of stating the alternative hypothesis is as follows:

$$H_1: \text{At least two of the } \mu_i \text{ are different.}$$

2.2 RATIONALE FOR ANOVA

Now that we know what the null hypothesis is that is being tested in a one way ANOVA, we might ask the following questions, “Why bother doing an ANOVA when comparing k groups? Why not simply do several t tests?” To see why the latter is problematic, let us consider the four group case ($T_1 T_2 T_3 T_4$). There are six paired comparisons here (12, 13, 14, 23, 24, 34). We could do six t tests, each at the .05 level, to determine which of these pairs are significantly different. Now, for just one of these t tests the α level is under control at .05. But, for the *set of 6* tests the α level gets out of control, since there is a 5% risk of false rejection for each test. We define the *overall α level* for a set of tests as the probability of at least one type I error (false rejection) when H_0 is true. Now, it can be shown that if α is small, then overall $\alpha \approx r\alpha$, where r is the number of tests being done. Actually, $r\alpha$ is an *upper bound* on the overall α level. Let us use this to see how rapidly the overall α inflates as the number of groups increases:

<i>Number of groups</i>	<i>Number of t tests</i>	<i>Approximate overall α</i>
3	3	.15
4	6	.30
5	10	.50
6	15	.75

This table shows that if we were to compare five groups with 10 t tests, each at the .05 level, then we have an approximate 50% chance of at least one false rejection. Thus, the probability of a few false rejections here is uncomfortably high. For six groups and 15 t tests, the probability of 2 or 3 false rejections is *very likely* (approximately .75)! Thus, it should be clear that using multiple t tests in a k group situation is not the way to proceed. We see later on in the chapter (Section 2.15) that a much tighter upper bound than $r\alpha$ can be put on overall α , especially when each test is done at the $\alpha = .05$ level.

2.3 NUMERICAL EXAMPLE

The analysis of variance procedure, which is appropriate, was developed by R. A. Fisher in an agricultural context back in the 1920s. ANOVA is based on the following three assumptions:

1. The observations are normally distributed on the dependent variable in each group.
2. The population variances for the groups are equal.
3. The observations in each group are independent.

We consider the effect of violations of these assumptions in detail in Section 2.7.

For our example, suppose a consumer organization wants to compare the price of a particular toy in three types of stores in a suburban county: variety stores, department stores, and discount toy stores. A random sample of 3 variety stores, 4 department stores, and 5 discount toy stores is selected and the following prices (in dollars) are recorded. We wish to test whether there is a difference in the average prices on this toy for the populations of stores from which these stores were selected.

The null hypothesis that is being tested here is

$$H_0: \mu_1 = \mu_2 = \mu_3$$

The sample means above are estimating the population means:

$$\bar{x}_1 = \hat{\mu}_1, \bar{x}_2 = \hat{\mu}_2, \bar{x}_3 = \hat{\mu}_3$$

<i>Variety</i>	<i>Dept.</i>	<i>Discount</i>
3	4	4
6	7	5
8	9	2
	8	3
		5
<hr/>		
$\bar{x}_1 = 5.67$	$\bar{x}_2 = 7$	$\bar{x}_3 = 3.8$

We wish to determine whether the sample means differ sufficiently, given sampling error, to suggest that the underlying population means differ. To determine this the ANOVA computes and compares two basic sources of variation:

1. Between group variation—determines how much the group means vary about the grand (overall) mean.
2. Within group variation—determines how much the subjects scores vary who are in the same group. Variation here is primarily due to individual differences.

Between Group Variation

The general formula here is given by

$$SS_b = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$$

$$SS_b = n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2 + \cdots + n_k (\bar{x}_k - \bar{x})^2 \quad (1)$$

where Σ is the summation symbol, n_i denotes the number of subjects in the i th group, \bar{x} denotes the grand mean, and SS_b stands for sum of squares between. It is a weighted sum of squares, where each deviation is weighted by the number of subjects in that group.

For the above data this becomes

$$\begin{aligned} SS_b &= \sum n_i (\bar{x}_i - 5.33)^2 \\ &= n_1 (\bar{x}_1 - 5.33)^2 + n_2 (\bar{x}_2 - 5.33)^2 + n_3 (\bar{x}_3 - 5.33)^2 \\ &= 3(5.67 - 5.33)^2 + 4(7 - 5.33)^2 + 5(3.8 - 5.33)^2 \\ &= .3468 + 11.1556 + 11.7045 = 23.2069 \end{aligned}$$

In calculating the grand (overall) mean above it is simplest to add up all the scores and divide by total number of subjects. Thus, in the above case this yields $\bar{x} = 64 / 12 = 5.33$. One can also obtain the grand mean from the individual means with the following formula:

$$\bar{x}_c = (n_1 \bar{x}_1 + n_2 \bar{x}_2 + \cdots + n_k \bar{x}_k) / N$$

where n_1 is the number of subjects in group 1, n_2 is the number of subjects in group 2, etc., and N represents total number of subjects. Note that this is a *weighted* average and that means based on a larger number of subjects receive greater weight in determining the grand mean. Because of this it is not appropriate to find the grand mean with unequal group sizes by simply taking the average of the means—a mistake frequently made.

We need the mean sum of square between (MS_b), since this represents a variance (we see why in Section 2.5). MS_b is simply sum of squares between (SS_b) divided by degrees of freedom, i.e.,

$$MS_b = SS_b / (k - 1) = 23.2069 / (3 - 1) = 11.6035 \quad (2)$$

Within Group Variation

Verbally, within group variability is calculated by deviating each score in group 1 about the mean in group 1, and squaring and summing these deviations. We then square the deviations of all scores of group 2 about the mean for group 2 and sum them, and so forth on to the k th group. We then pool (add) these squared deviations to obtain the sum of squares within, denoted by SS_w . Symbolically then, it looks like

$$SS_w = \sum_1 (x_{i1} - \bar{x}_1)^2 + \sum_2 (x_{i2} - \bar{x}_2)^2 + \cdots + \sum_k (x_{ik} - \bar{x}_k)^2 \quad (3)$$

where x_{i1} is the score of the i th subject in group 1, x_{i2} is the score of the i th subject in group 2, and x_{ik} is the score of the i th subject in group k . Now we calculate SS_w for the above data:

$$\begin{aligned} SS_w = & (3 - 5.67)^2 + (6 - 5.67)^2 + (8 - 5.67)^2 \text{ (variability within gp 1)} \\ & + (4 - 7)^2 + (9 - 7)^2 + (8 - 7)^2 \text{ (variability within gp 2)} \\ & + (4 - 3.8)^2 + (5 - 3.8)^2 + (2 - 3.8)^2 + (3 - 3.8)^2 + (5 - 3.8)^2 \\ & \text{(variability within gp 3)} \end{aligned}$$

$SS_w = 33.4667$ pooled within group variability. Once again we need MS_w (mean sum of squares within), which represents a variance, rather than SS_w . The formula for MS_w is

$$MS_w = SS_w / (N - k) = 33.4667 / (12 - 3) = 3.7185 \quad (4)$$

where N denotes the total number of subjects.

The F Test

To test the tenability of the null hypothesis the following F statistic is used: $F = MS_b / MS_w$. Thus, for our data this is

$$F = MS_b / MS_w = 11.6035 / 3.7185 = 3.12 \quad (5)$$

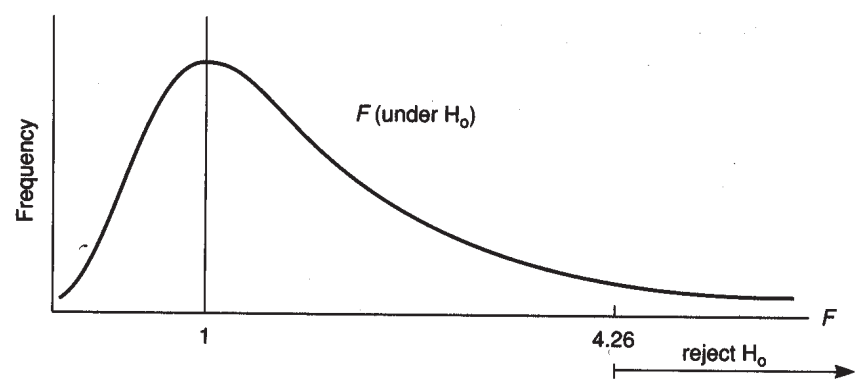
To determine whether this is large enough to reject H_0 we must ascertain if this is a very unlikely value to occur if indeed the null hypothesis is true. To get at this we need to refer to the *sampling distribution of F under H_0 (assuming the population means are equal)*. Here we must think conceptually as follows. If we were to draw samples of sizes 3, 4, and 5 repeatedly from populations with equal means, and compute an F ratio for each draw, what would the distribution of F 's look like? This is the sampling distribution of F under H_0 . Statisticians have determined that the distribution will be positively skewed, with a modal value of approximately 1. We will sketch the distribution shortly.

Now, suppose we are testing H_0 at $\alpha = .05$. The above sampling distribution will have the following pair of degrees of freedom:

$$\begin{aligned} df_b &= k - 1 \text{ (degrees of freedom between) and} \\ df_w &= N - k \text{ (degrees of freedom within)} \end{aligned}$$

Thus, for our data $df_b = 3 - 1 = 2$ and $df_w = 12 - 3 = 9$. For an F distribution with 2 and 9 degrees of freedom it has been determined that the 95th percentile point is 4.26 (the point corresponding to testing at .05 level). That is, if the null hypothesis is true, then only 5% of the time would we expect to obtain an F greater than 4.26. Thus, 4.26 is our critical value, and if the value of the test statistic is greater than 4.26 we will reject H_0 . For our case $F = 3.12$, so we fail to reject H_0 and conclude

that it is possible that the population means are equal. The sampling distribution is sketched below:



The results of an ANOVA are typically summarized in a table as follows:

Source	SS	Df	MS	F
Between	$\Sigma n_i(\bar{x}_i - \bar{x})^2$	$k - 1$	$SS_b / (k - 1)$	$MS_b \setminus MS_w$
Within	$\Sigma(\bar{x}_{i1} - \bar{x}_1)^2 + \Sigma(\bar{x}_{i2} - \bar{x}_2)^2 + \dots + \Sigma(\bar{x}_{ik} - \bar{x}_k)^2$	$N - 1$	$SS_w / (N - k)$	

For the above example this table would be:

Source	SS	df	MS	F
Between	23.2069	2	11.6035	3.12
Within	33.4667	9	3.7185	

The critical values for ANOVAs with varying sample size and different α levels have been tabled, and are found in Table B.1 at the end of this book. To give the reader practice in using these tables we consider two examples.

Example 1

An experimenter runs a 3 group ANOVA with 10 subjects per group, testing at $\alpha = .10$. He obtains $F = 2.16$. Does he reject the null hypothesis? First we find degrees of freedom between and within: $df_b = 3 - 1 = 2$ and $df_w = 30 - 3 = 27$. Reference to Table B.1 then shows that the critical value $= F_{.10;2,27} = 2.51$, and thus he would fail to reject the null hypothesis.

Example 2

An investigator runs a 4 group ANOVA with following sample sizes: $n_1 = 15$, $n_2 = 20$, $n_3 = 10$, and $n_4 = 25$. She will test at $\alpha = .01$. Will she reject H_0 if she obtains $F = 5.26$? The degrees of freedom are $df_b = 3$ and $df_w = 70 - 4 = 66$. Reference to Table B.1 shows that critical value for 3 and 66 degrees of freedom is not in the table. What do we do here? Note that the tabled values for error degrees of freedom are given from 1 to 30 and then jump to 40, 60, 120, and infinity. The reason is that the critical values change very little once the degrees of freedom gets beyond 30. We could interpolate in our case between 60 and 90, but since the values change so little our recommendation is simply to use the critical value for the closest error degrees of freedom, which here is 60. Thus, the critical value is $F_{.01;3,60} = 4.13$. Since the value of $F = 5.26$, which is greater than 4.13, we reject the null hypothesis.

When we reject the null hypothesis at some α level all we know is that there is an *overall difference* among the groups. To locate where the differences lie (e.g., which pairs of groups are significantly different) we need some post hoc (after this—from the Latin) procedure. Many such post hoc procedures have been developed, and we consider these in Section 2.10.

2.4 EXPECTED MEAN SQUARES

Earlier we stated that the modal value of F in the sampling distribution under H_0 was about 1; i.e., this is the value we would *expect* to obtain most frequently *if* indeed the population means are equal. In other words, we were saying that the expected value of F is about 1, or in symbols $E(F) \approx 1$. The reason this is true is because of the expected values for MS_b and MS_w under the null hypothesis.

The reader can think of expected value as the long term average. As a simple example, consider flipping a coin 1,000 times. Then, if it is a fair coin, we would expect about 500 heads and 500 tails, that is, $E(H) = E(T) = .5$. In the ANOVA context we think of repeating the experiment thousands of times, and computing a MS_b and MS_w each time. Now, it can be shown that if we were to average the thousands of MS_w 's, then the average would be σ^2 . Recall that σ^2 was the assumed common population variance for the groups. That is, σ^2 involved one of the assumptions underlying the analysis of variance. Thus we have

$$E(MS_w) = \sigma^2 \text{ (when } H_0 \text{ is true)} \quad (6)$$

Also, it is important to note that the size of MS_w does *not* depend on whether the population means differ or not. However, we will see that the $E(MS_b)$ does depend on differences in population means. It can be shown that

$$E(MS_b) = \sigma^2 + n_i \sum (\mu_1 - \mu)^2 / (k - 1) \quad (7)$$

If the population means are equal (H_0 is true), then this means $\mu_1 = \mu_2 = \dots = \mu_k = \mu$ and the second term in the above expression will be 0, or in other words,

$$E(MS_b) = \sigma^2 \text{ (when } H_0 \text{ is true)} \quad (8)$$

Thus, when the null hypothesis is true, the expected values for numerator and denominator of the F ratio are equal and the expected value of F statistic is equal to about 1. The reason that $E(F)$ is not exactly equal to 1 is because the expected value of a quotient is not equal to the quotient of expected values.

Thus, evidence in favor of rejecting H_0 will be reflected in an F ratio greater than 1. How much greater than 1 the F must be to reject H_0 depends a great deal on sample size. The next chapter on power deals with this issue in detail.

2.5 MS_w AND MS_b AS VARIANCES

We mentioned earlier that MS_w and MS_b actually represent variances. Now we show this algebraically. This is easiest to see for equal n per group and so that is demonstrated. Using Equations 3 and 4, we can write MS_w as

$$MS_w = \left[\sum (x_{i1} - \bar{x}_1)^2 + \sum (x_{i2} - \bar{x}_2)^2 + \dots + \sum (x_{ik} - \bar{x}_k)^2 \right] / (N - k)$$

For equal n per group, we have $N = nk$ and therefore we have $N - k = nk - k = k(n - 1)$. Thus, we can rewrite the above equation as

$$MS_w = 1 / k(n - 1) \left[\sum (x_{i1} - \bar{x}_1)^2 + \sum (x_{i2} - \bar{x}_2)^2 + \dots + \sum (x_{ik} - \bar{x}_k)^2 \right]$$

or

$$MS_w = 1 / k \left[\underbrace{\frac{\sum (x_{i1} - \bar{x}_1)^2}{n - 1}}_{\text{variance for gp 1}} + \underbrace{\frac{\sum (x_{i2} - \bar{x}_2)^2}{n - 1}}_{\text{variance for gp 2}} + \dots + \underbrace{\frac{\sum (x_{ik} - \bar{x}_k)^2}{n - 1}}_{\text{variance for gp } k} \right]$$

Recall from beginning statistics that each is in the *form* of a variance, since variance for a single group of subjects is $s^2 = \sum (x_i - \bar{x})^2 / (n - 1)$.

Thus, for equal group size, MS_w is just the average of the sample variances for the groups.

Using Equations 1 and 2, we can write MS_b as follows:

$$MS_b = \left[n(\bar{x}_1 - \bar{x})^2 + n(\bar{x}_2 - \bar{x})^2 + \dots + n(\bar{x}_k - \bar{x})^2 \right] / (k - 1)$$

or

$$MS_b = n \underbrace{\sum (\bar{x}_1 - \bar{x})^2 / (k-1)}_{\text{variance for } k \text{ group means about the grand mean}}$$

Thus, MS_b is a *weighted variance of the group means about the grand mean*. This is somewhat more subtle, but note that except for the n we have the *form of a variance*, where the group means are playing the role of individual observations and the grand mean is playing the role of the mean for a single group.

2.6 A LINEAR MODEL FOR THE DATA

We now state the *linear* model for each subject's score on which the one way ANOVA is based. The model for the score of the i th subject in group j (y_{ij}) is given by:

$$y_{ij} = \mu + \alpha_j + e_{ij} \quad (9)$$

where μ is the grand mean for all subjects, $\alpha_j = \mu_j - \mu$ is the treatment effect for the j th treatment, and e_{ij} is the random error for the i th subject in the j th treatment.

Thus, we are postulating that a subject's score is composed of three parts: (1) a general effect—the grand mean, (2) an effect unique and constant within a given treatment (α_j), and (3) an effect that is unpredictable, that is, e_{ij} . It is assumed that the e_{ij} are independent, normally distributed within each treatment, and have the *same* variance for each treatment. Note that these assumptions for the e_{ij} imply exactly the same assumptions for the y_{ij} (the subjects' scores), since e_{ij} is the only random part of the model on the right side of Equation 9.

To gain some feeling for the above linear model, consider three treatments with 4 subjects in each treatment. Suppose there is just a general effect (i.e., no treatment effect or random error). Then the data would look like this:

MODEL: $y_{ij} = \mu$

T_1	T_2	T_3
20	20	20
20	20	20
20	20	20
20	20	20

Next, suppose there is in addition a treatment effect but no random error. Then the data might look like this:

MODEL: $y_{ij} = \mu + \alpha_j$,

T_1	T_2	T_3
19	17	24
19	17	24
19	17	24
19	17	24

In the above we have $\alpha_1 = -1$, $\alpha_2 = -3$, and $\alpha_3 = 4$.

But both of the above situations are too simple for real data, since subjects' scores will essentially always vary within each treatment group. The main reason they will differ is because of individual differences (they come to the treatments with different capabilities, backgrounds, motivation, etc.). Measurement error also contributes to within treatment variability. Since the y_{ij} will vary, this implies the random error components will vary. If we now add an error component to each subject's score we might obtain a realistic data set like this:

$$\text{MODEL: } y_{ij} = \mu + \alpha_j + e_{ij}$$

T_1	T_2	T_3
18	18	23
24	19	22
21	11	27
17	16	28

Here we have added an error component of $e_{11} = -1$ for the first subject's score in treatment 1, an $e_{21} = 5$ for the second subject in treatment 1, an $e_{31} = 2$ for the third subject, etc.

2.7 ASSUMPTIONS IN ANOVA

We mentioned earlier that the analysis of variance is based on the following three assumptions:

1. The observations are normally distributed on the dependent variable in *each* group.
2. The population variances for the groups are equal. This is the so-called *homogeneity of variance* assumption. In symbols this would be $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2 = \sigma^2$
3. The observations are independent.

Why is it important to study the assumptions underlying ANOVA? Because in ANOVA we set up a mathematical model based on the assumptions, and all mathe-

mathematical models are approximations to reality. Therefore, violations of the assumptions are inevitable. The salient question becomes, "How radically must a given assumption be violated before it has a serious effect on type I and type II error rates?" Thus, we may set our $\alpha = .05$ and think we are rejecting falsely 5% of the time, but if a given assumption is violated we may be rejecting falsely 40% of the time. For these kind of situations we would certainly want to be able to detect such violations and take some corrective action. But all violations of assumptions are not serious, and hence it is crucial to know *which* assumptions to be particularly concerned about, and under what conditions. Before we begin our review of a considerable literature on violations of assumptions in ANOVA, it is helpful to cover some basic terminology that is needed in discussing the results of Monte Carlo (i.e., simulation) studies.

The nominal α (level of significance) is the level set by the experimenter, and is the percent of time one is rejecting falsely when the null hypothesis is true *and all* assumptions are met. The actual α is the percent of time one is rejecting falsely if one or more of the assumptions is violated. We say a test statistic is *robust* if the actual α is very close to the nominal α .

Numerous studies have examined the effect of violations of assumptions in ANOVA, and an excellent summary of this literature has been provided by Glass, Peckham, and Sanders (1972). Their review indicates that non-normality has only a slight effect on the type I error rate, even for very skewed or kurtotic distributions. For example, the actual α s for some very non-normal populations were only .055 or .06: very minor deviations from the nominal level of .05. We say the F statistic is robust with respect to the normality assumption.

The reader may be puzzled as to how this can be. The basic reason is the *Central Limit Theorem*, which states that the sum of independent observations having any distribution whatsoever approaches a normal distribution as the number of observations increases. To be somewhat more specific, Bock (1975) notes, "even for distributions which depart markedly from normality, sums of 50 or more observations approximate to normality. For moderately non-normal distributions the approximation is good with as few as 10 to 20 observations" (p. 111). Now, since the sums of independent observations approach normality rapidly, so do the means, and the sampling distribution of F is based on means. Thus, the sampling distribution of F is only slightly affected, and therefore the critical values when sampling from normal and non-normal distributions will not differ by much.

Lack of normality due to skewness also has only a slight effect on power (a few hundredths). Platykurtosis (a flattened out distribution relative to the normal) does affect power, and the effect can be substantial for small n .

Now, we deal with the second assumption, homogeneity of the population variances. If the group sizes are equal or approximately equal (largest/smallest < 1.5), then the F statistic is robust for unequal variances. That is, the actual α stays close to the nominal α (level of significance). The only time one need worry is when the

group sizes are sharply unequal (largest/smallest > 1.5) *and* a statistical test shows that the population variances are unequal. For this class of situations the studies have found that if the large variances are associated with the small group sizes, then F is *liberal*. A statistic being liberal means we are rejecting falsely too often, i.e., the actual $\alpha >$ nominal α . Thus, an experimenter may think he or she is rejecting falsely 5% of the time (nominal α), but in fact the true rejection rate may be 11% (actual α). On the other hand, when the large variances are associated with the large group sizes, then the F statistic is *conservative*. This means the actual $\alpha <$ nominal α . Many researchers would not consider this serious; however, note that the smaller α will cause a decrease in power.

There are many statistical tests for homogeneity of variance (e.g., Bartlett's, Cochran's, Hartley's F_{\max}), but these all suffer from being very sensitive to non-normality. That is, one may reject with these tests and conclude that the population variances are different when in fact the rejection may have been due to non-normality in the underlying populations. Fortunately there is a test, due to Levene, which is somewhat more robust against non-normality, and it is available on SAS and SPSS.

Examples

Consider the three data situations below. In which (if any) of the cases would you be concerned?

	Case 1			Case 2			Case 3			
	GROUPS			GROUPS			GROUPS			
	1	2	3	1	2	3	1	2	3	4
n_i	18	20	17	20	50	30	10	30	15	25
s_i^2	15	90	42	80	10	35	50	100	70	140

In Case 1 there is no need to be concerned since the group cases are approximately equal ($20/17 < 1.5$), and therefore F is robust. In Case 2 the group sizes are sharply unequal ($50/20 > 1.5$), so there will be a problem if a statistical test shows the population variances to be different. We use Hartley's F_{\max} = largest variance/smallest variance, assuming normality is not a problem, and find that $F_{\max} = 80/10 = 8$. Referring to Table B.4, and using the average group size (35) to enter the table, we find that the critical value at .05 is about 2.4. Thus, we conclude the population variances are different. In this case the large sample variances are associated with the smaller group sizes, so that F will be liberal. What is to be done here? There are at least 3 possibilities. One is to do an ANOVA which does not assume equal variances. Another choice is to simply test at a more stringent α level (say

.01), realizing that the actual α will probably be in the vicinity of .05. A third choice is to seek help from a statistician on a variance stabilizing transformation (such as square root or log). I would recommend either of the first two choices for applied researchers.

In Case 3 we again have a potential problem because the group sizes are sharply unequal ($30/10 > 1.5$). Using Table B.4 and an average group size of 20 to enter the table, we find $F_{\max} = 140/50 = 2.8$. This is not significant since the critical value = 3.29. Therefore there is no problem here since the assumption is tenable.

So far we have treated heterogeneity of variance as a nuisance, something we wish will not happen so that the analysis on the means can proceed accurately. However, unequal variances, or a focus on dispersion, in some situations may be an interesting and important finding. Raudenbusch and Bryk (1987) cite a study by Bryk (1977) in which a compensatory program that increased mean achievement also increased dispersion in achievement. As they note, such a program might increase the number of children failing to attain some minimum standard even though it raised mean achievement. Also, occasionally variance reduction is an explicit goal of educational programs, as in some mastery learning programs (Bloom, 1984).

2.8 THE INDEPENDENCE ASSUMPTION

Although we have listed the independence assumption last, it is *by far the most important assumption, for even a small violation of it produces a substantial effect on both the level of significance and the power of the F statistic*. Just a small amount of dependence among the observations causes the actual α to be several times greater than the nominal α . Dependence among the observations is measured by the intraclass correlation R , where:

$$R = (MS_b - MS_w) / (MS_b + (n - 1)MS_w) \quad (10)$$

MS_b and MS_w are the numerator and denominator of the F statistic and n is the number of subjects per group.

Table 2.1, from Scariano and Davenport (1987), shows precisely how dramatic of an effect dependence has on type I error. For example, for the 3 group case with 10 subjects per group and moderate dependence (intraclass correlation = .30) the actual α is .5379! Also, for 3 groups with 30 subjects per group and small dependence (intraclass correlation = .10) the actual α is .4917, almost 10 times the nominal α of .05! Notice, also from the table that for a fixed value of the intraclass correlation the situation does not improve with larger sample size, but gets far worse.

Now let us consider some situations in social science research where dependence among the observations will be present. Teaching methods studies constitute a broad class of situations where dependence is undoubtedly present. For example,

a few troublemakers in a classroom would have a detrimental effect on the achievement of many children in the classroom. Thus, their posttest achievement would be at least partially dependent on the disruptive classroom atmosphere. On the other hand, even in a good classroom atmosphere, dependence is introduced, for the achievement of many of the children will be enhanced by the positive learning situation. Therefore, in either case (positive or negative classroom atmosphere), the achievement of the children is not independent of the other children in the classroom.

Another situation I came across recently in which dependence among the observations was present involved a study comparing the achievement of students working in pairs at microcomputers vs. students working in groups of three at the micros. Here, if Bill and John are working at the same microcomputer, then obviously Bill's achievement is partially influenced by John. The proper unit of analysis in this study is the *mean* achievement for each pair and triplet of students, as it is plausible to assume that the achievement of students on one micro is independent of the students working at the other micros.

Glass and Hopkins (1984) make the following statement concerning situations where independence may or may not be tenable: "Whenever the treatment is individually administered, observations are independent. But where treatments involve interaction among persons, such as 'discussion' method or group counseling, the observations may influence each other" (p. 353).

What Should Be Done With Correlated Observations?

Given the results in Table 2.1 for a positive intraclass correlation, one route investigators should seriously consider if they suspect that the nature of their study will lead to correlated observations is to test at a more stringent level of significance. For the 3 and 5 group cases in Table 2.1 with 10 observations per group and intraclass correlation = .10, the error rates are 5 to 6 times greater than the assumed level of significance of .05. Thus, for this type of situation it would be wise to test at $\alpha = .01$, realizing that your actual error rate will be about .05 or somewhat greater. For the 3 and 5 group cases in Table 2.1 with 30 observations per group and intraclass correlation = .10, the error rates are about *10 times* greater than .05. Here, it would be advisable to either test at $\alpha = .01$, realizing that actual α will be about .10, or test at an even more stringent level than .01.

If several small groups (counseling, social interaction, etc.) are involved in each treatment, and there is clear reason to suspect that subjects' observations will be correlated within groups, but that observations will not be correlated across the groups, then consider using the *group mean as the unit of analysis*. Of course this will reduce the effective sample size considerably; however, this will not cause as drastic a drop in power as some have feared. The reason is that the means are much

TABLE 2.1
Actual Type I Error Rates for Correlated Observations in a One Way
ANOVA (Nominal $\alpha = .05$)

<i>m</i>	<i>n</i>	<i>Intraclass Correlation</i>								
		.00	.01	.10	.30	.50	.70	.90	.95	.99
2	3	.0500	.0522	.0740	.1402	.2374	.3819	.6275	.7339	.8800
	10	.0500	.0606	.1654	.3729	.5344	.6752	.8282	.8809	.9475
	30	.0500	.0848	.3402	.5928	.7205	.8131	.9036	.9335	.9708
	100	.0500	.1658	.5716	.7662	.8446	.8976	.9477	.9640	.9842
3	3	.0500	.0529	.0837	.1866	.3430	.5585	.8367	.9163	.9829
	10	.0500	.0641	.2227	.5379	.7397	.8718	.9639	.9826	.9966
	30	.0500	.0985	.4917	.7999	.9049	.9573	.9886	.9946	.9990
	100	.0500	.2236	.7791	.9333	.9705	.9872	.9966	.9984	.9997
5	3	.0500	.0540	.0997	.2684	.5149	.7808	.9704	.9923	.9997
	10	.0500	.0692	.3151	.7446	.9175	.9798	.9984	.9996	1.0000
	30	.0500	.1192	.6908	.9506	.9888	.9977	.9998	1.0000	1.0000
	100	.0500	.3147	.9397	.9945	.9989	.9998	1.0000	1.0000	1.0000
10	3	.0500	.0560	.1323	.4396	.7837	.9664	.9997	1.0000	1.0000
	10	.0500	.0783	.4945	.9439	.9957	.9998	1.0000	1.0000	1.0000
	30	.0500	.1594	.9119	.9986	1.0000	1.0000	1.0000	1.0000	1.0000
	100	.0500	.4892	.9978	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

m—number of groups

n—number of observations per group

more stable than individual observations and hence the within variability will be far less.

Table 2.2, from Barcikowski (1981), shows that if the effect size is medium or large, then the number of groups needed per treatment for power $> .80$ doesn't have to be that large. For example, at $\alpha = .10$, intraclass correlation = .10, and medium effect size, 10 groups (of 10 subjects each) are needed per treatment. For power $> .70$ (which I consider adequate) at $\alpha = .15$ one probably could get by with about 5 or 6 groups of 10 per treatment. This is a rough estimate, since it involves double extrapolation.

Before we leave the topic of correlated observations, we wish to mention an interesting paper by Kenny and Judd (1986), who discuss how non-independent observations can arise because of several factors, grouping being one of them. The following quote from their paper is important to keep in mind for applied researchers:

Throughout this article we have treated nonindependence as a statistical nuisance, to be avoided because of the bias that it introduces.... There are, however, many occasions when nonindependence is the substantive problem that we are trying to under-

TABLE 2.2
Number of Groups per Treatment Necessary for Power > .80 in a Two
Treatment Level Design

α level	<i>Effect Size</i>		<i>Intraclass Correlation</i>				
	<i>Number per group</i>	<i>.10</i>			<i>.20</i>		
		<i>.20</i>	<i>.50</i>	<i>.80</i>	<i>.20</i>	<i>.50</i>	<i>.80^a</i>
.05	10	73	13	6	107	18	8
	15	62	11	5	97	17	8
	20	56	10	5	92	16	7
	25	53	10	5	89	16	7
	30	51	9	5	87	15	7
	35	49	9	5	86	15	7
	40	48	9	5	85	15	7
	10	57	10	5	83	14	7
.10	15	48	9	4	76	13	6
	20	44	8	4	72	13	6
	25	41	8	4	69	12	6
	30	39	7	4	68	12	6
	35	38	7	4	67	12	5
	40	37	7	4	66	12	5

^a.20—small effect size
.50—medium effect size
.80—large effect size

stand in psychological research. For instance, in developmental psychology, a frequently asked question concerns the development of social interaction. Developmental researchers study the content and rate of vocalization from infants for cues about the onset of interaction. Social interaction implies nonindependence between the vocalizations of interacting individuals. To study interaction developmentally, then, we should be interested in nonindependence not solely as a statistical problem but also a substantive focus in itself. ... In social psychology, one of the fundamental questions concerns how individual behavior is modified by group contexts. (p. 431)

2.9 ANOVA ON SPSS AND SAS

Now we consider how to run a one way ANOVA on SPSS and SAS, and how to interpret the printout. To illustrate we shall use the following 4 group data set:

<i>Group 1</i>	<i>Group 2</i>	<i>Group 3</i>	<i>Group 4</i>
2	7	4	8
3	9	4	4
5	11	5	7
6		8	7
		3	

The complete SAS control lines for obtaining the ANOVA and Tukey procedure are presented in Table 2.3, and the annotated SAS printout is given in Table 2.4. We also ran the above data on SPSS for Windows 12.0, and appropriate screens are given in Table 2.5. First click on ANALYZE, then scroll down to COMPARE MEANS and over to ONE WAY ANOVA. When you click on ONE WAY ANOVA and select Y as the dependent variable and GPID as the factor, the screen appears as in the middle of Table 2.5. When you select POST HOC and select TUKEY, the screen appears as in the bottom of Table 2.5. Selected ANOVA and Tukey printout from SPSS for Windows 12.0 appears in Table 2.6.

TABLE 2.3
SAS Control Lines for One Way ANOVA and Tukey Procedure
on Sample Problem

	DATA INTERM;
①	INPUT GPID Y @@;
	LINES;
②	1 2 1 3 1 5 1 6
	2 7 2 9 2 11
	3 4 3 4 3 5 3 8 3 3
	4 8 4 4 4 7 4 7
③	PROC MEANS;
	BY GPID;
	PROC ANOVA;
④	CLASS GPID;
	MODEL Y = GPID;
	MEANS GPID/TUKEY;
⑤	PROC PRINT;

① There is a semicolon at the end of every SAS command, except for the data lines. The @@ is needed in order to put data for more than one subject on the same line.

② The first number of each pair is the group identification of the subject and the second number is the score on the dependent variable.

③ This PROC MEANS is necessary to obtain the means on the dependent variable in each group.

④ The ANOVA procedure is called and GPID is identified as the grouping (independent) variable through this CLASS statement.

⑤ This procedure provides a listing of the data.

TABLE 2.4
Selected Printout from SAS ANOVA for a One Way ANOVA

Insert art (INCLUDING FOOTNOTES) from p. 82 of previous edition

TABLE 2.5
SPSS for Windows 12.0 Screens for One Way ANOVA and Tukey
Procedure on Sample Problem

Insert art from p. 83 of previous edition

TABLE 2.6
Selected ANOVA and Tukey Procedure Printout From SPSS for Windows
12.0

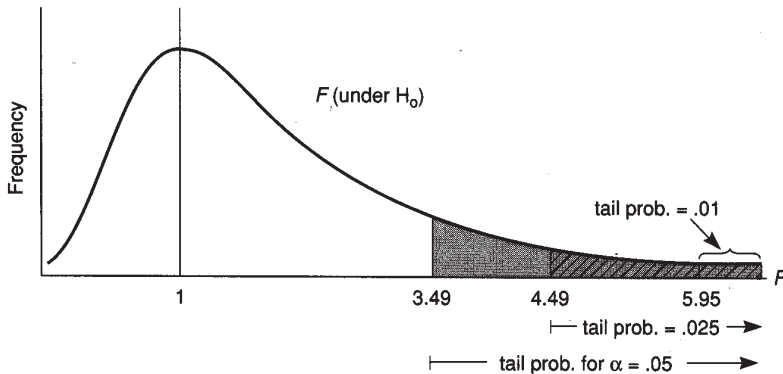
Insert art from p. 84 of previous edition

p Values (Tail Probabilities)

In Tables 2.4 and 2.6 the reader will note to the right of the F statistic for the ANOVA the following for SAS ($PR > F$) and for SPSS (SIG), with a numerical value of .0196 in both cases. Although labeled somewhat differently by the packages, these are p values, or tail probabilities. It is the probability of obtaining an F larger than 4.857 when the null hypothesis is true (population means are equal). If we had set $\alpha = .05$ a priori, then we would reject H_0 here since we are willing to take a 5% chance of rejecting falsely, and the tail probability indicates there is only about a 2% chance. These tail probabilities, which are printed out on all the major statistical packages, eliminate the need to look up critical values. We can adopt the following rules:

tail prob. $< \alpha$ level \Rightarrow reject at that α level

tail prob. $> \alpha$ level \Rightarrow fail to reject at that α level



Students often get confused when told to reject if the tail probability is *less* than the α level, since they may have been told repeatedly in an introductory statistics course to reject if the value of the test statistic is *greater* than the critical value. The connection that needs to be made here is to see that if the tail probability is $< \alpha$ level, then the value of the test statistic must be in the critical region. To illustrate this, suppose that in the computer problem we had tested F for significance by using the critical value at the .05 level, which is 3.49. It is very important to note that the critical value cuts off a tail probability of .05, i.e., only 5% of F values will be greater than 3.49 if H_0 is true (this area is shaded in the diagram below). Now, any value of F greater than 3.49 (in the critical region) must have a tail probability *less* than .05. Why? Because any $F > 3.49$ will have a smaller area under the F distribution, and this area represents the tail probability. For example, $F = 4.49$ has a tail

probability = .025; the lined area, while $F = 5.95$ has a tail probability = .01. Notice that both of these F s (4.49 and 5.95) are in the critical region.

Huberty (1987) has written an interesting article in which he discusses p values and notes:

The lack of discussion in textbooks written for behavioral science researchers is somewhat puzzling in light of the common practice of reporting P -values (the lower-case p is often used) in journal articles, and in light of attention paid to them in publication manuals (e.g., American Psychological Association), (p. 5)

2.10 POST HOC PROCEDURES

As mentioned earlier, there are numerous post hoc procedures available for determining where the differences lie after the F statistic has indicated there is a significant overall difference. Among the post hoc procedures are the Tukey, Scheffé, Newman-Keuls, Duncan, and Fisher's LSD (the so-called protected t test). All these procedures have two fundamental purposes:

1. To enable us to ferret out where the differences lie, and
2. To maintain the overall a level (or experimentwise error rate) at some pre-determined level, usually set at .05. In other words, keep a lid on the probability of false rejections for all the tests being done.

Unfortunately, the Newman-Keuls, Duncan, and Fisher's LSD do not control overall α as claimed. That is, they tend to be liberal. On the other hand, the Scheffé procedure tends to be quite conservative (since it allows for a wide range of comparisons to be done). We discuss and illustrate the Scheffé procedure in Section 2.12. For paired comparisons we favor and present the Tukey procedure for 3 reasons:

1. The Tukey *does* control the overall α as claimed (Hayter, 1984).
2. The Tukey procedure examines a focused, meaningful, and easily interpreted set of comparisons, that is, all paired comparisons.
3. The Tukey is a fairly powerful procedure for detecting differences.

Thus the Tukey provides a nice balance in terms of controlling on both type I and type II errors, while focusing on meaningful, easily interpreted comparisons.

However, if you are *only* interested in comparing each of several treatment groups against a control group, then the Dunnett (1955) procedure is most powerful and should be used (see Exercise 16).

2.11 TUKEY PROCEDURE

The Tukey procedure, which is sometimes called the HSD (honestly significant difference) test, enables us to examine *all pairwise* group comparisons with the experimentwise (overall) α level held in check. The studentized range statistic (which we denote by q) is used in the procedure, and the critical values for it are given in Table B.2 in the back of the book. The procedure establishes a set of *simultaneous confidence intervals* for each pair of population means. The intervals are given by:

$$(\bar{x}_i - \bar{x}_j) \pm q_{\alpha; k, N-k} \sqrt{MS_w / n} \quad (11)$$

or

$$(\bar{x}_i - \bar{x}_j) - q_{\alpha; k, N-k} \sqrt{MS_w / n} < \mu_i - \mu_j < (\bar{x}_i - \bar{x}_j) + q_{\alpha; k, N-k} \sqrt{MS_w / n}$$

where \bar{x}_i and \bar{x}_j represent the means for any two groups, q is just a tabled value, MS_w is the denominator of the F statistic, and n is the assumed common group size.

In deriving the procedure, Tukey assumed equal group sizes. In practice, however, often the group sizes are not equal. Does this severely limit the utility of the procedure? No, since various studies (Dunnett, 1980; Kesselman, Murray, & Rogan, 1976) indicate that the Tukey still controls overall α *provided that the population variances are equal and that n is replaced by the harmonic mean $2n_1n_2/(n_1 + n_2)$ for each pair of groups.* The harmonic mean for each pair of groups is what is used by default for both SAS and SPSS. Thus, for unequal group sizes the n in Equation 11 is replaced by $2n_i n_j / (n_i + n_j)$ when comparing groups i and j . When this replacement is made, it is called the Tukey-Kramer procedure. Now let us consider a numerical example to illustrate how to calculate and interpret the intervals.

Example

Consider the following 4 group problem with unequal group sizes. The reader may check with Hartley's F_{\max} that homogeneity of variance is tenable here.

	1	2	3	4
n_i	20	17	14	18
\bar{x}_i	7	8	10	13
s_i	4	5	6	4

A one way ANOVA on this data yields $F = 129.02/22.22 = 5.81$ ($p < .05$). Therefore we know there is a significant overall difference among the groups. To locate the pairs that are significantly different we use the Tukey procedure, with overall $\alpha = .05$. First, we need the harmonic means for each pair of groups:

Groups	Harmonic mean
1 and 2	$2(20)(17)/37 = 18.38$
1 and 3	$2(20)(14)/34 = 16.47$
1 and 4	$2(20)(18)/38 = 18.95$
2 and 3	$2(17)(14)/31 = 15.35$
2 and 4	$2(17)(18)/35 = 17.49$
3 and 4	$2(14)(18)/32 = 15.75$

The tabled value is $q_{.05;4,65} = 3.74$. Now we set up the intervals:

Differences	Critical value	Confidence intervals
$\bar{x}_1 - \bar{x}_2 = -1$	$3.74\sqrt{22.22 / 18.38} = 4.11$	$(-5.11, 3.11)$
$\bar{x}_1 - \bar{x}_3 = -3$	$3.74\sqrt{22.22 / 16.47} = 4.34$	$(-7.34, 1.34)$
$\bar{x}_1 - \bar{x}_4 = -6$	$3.74\sqrt{22.22 / 18.95} = 4.05$	$(-10.05, -1.95)$
$\bar{x}_2 - \bar{x}_3 = -2$	$3.74\sqrt{22.22 / 15.35} = 4.50$	$(-6.5, 2.5)$
$\bar{x}_2 - \bar{x}_4 = -5$	$3.74\sqrt{22.22 / 17.49} = 4.22$	$(-9.22, -.78)$
$\bar{x}_3 - \bar{x}_4 = -3$	$3.74\sqrt{22.22 / 15.75} = 4.44$	$(-7.44, 1.44)$

Note that the lower limit for the first interval is obtained by subtracting the critical value (4.11) from the difference in the means (-1) and the upper limit is found by adding the critical value to -1. Therefore

$$-1 - 4.11 < \mu_1 - \mu_2 < -1 + 4.11 \text{ or } -5.11 < \mu_1 - \mu_2 < 3.11$$

How do we interpret these intervals? First, if the confidence interval *includes 0* we conclude the population means are not different. Why? Because if the interval includes 0 it means 0 is a possible value for $\mu_i - \mu_j$, which is to say it is possible that $\mu_i - \mu_j$. Thus, in comparing groups 1 and 2 above, we see that the interval for $\mu_1 - \mu_2$ is given by

$$-5.11 < \mu_1 - \mu_2 < 3.11$$

Therefore, 0 is a possible value for $\mu_1 - \mu_2$, since 0 is in the interval. Thus, groups 1 and 2 are not significantly different. On the other hand, if the interval does not include 0, then the groups are significantly different, since 0 is not a possible value for the population mean difference. Examining the above intervals, we find that only groups 1 and 4, and groups 2 and 4 are significantly different.

Confidence intervals are more informative than tests of significance because they both indicate significance *and* give a range of values within which the population mean difference probably lies. Thus, confidence intervals are one way of judging the practical significance of results. Consider groups 1 and 4. Suppose a researcher had decided a priori that the population mean difference had to be at least

4 units to be of any practical significance. Now, the confidence interval for groups 1 and 4 is:

$$-10.05 < \mu_1 - \mu_4 < -1.95$$

and the result would not be practically significant because the difference could be as small as -1.95 .

2.12 THE SCHEFFÉ PROCEDURE

The big advantage of the Scheffé procedure is its flexibility. For a k group problem one can examine *all possible simple (pairwise) and complex* contrasts (this is defined very shortly) among the group means with the assurance that the overall α will be less than some preassigned value (say .05). Thus in exploratory research this procedure provides the ultimate in data snooping potential. However, in order to keep overall $\alpha = .05$, while doing all the statistical tests for the very large number of comparisons possible, the critical value necessary for significance will be large relative to what it would be for other multiple comparison procedures. This means power will suffer, which will not be of concern if sample size in your study is large (say about 100 subjects per group). If your group sizes are small (about 20 subjects per group), however, then on power considerations it would be wise to set overall α at .10, or even at .15.

In general, for k groups with population means $\mu_1, \mu_2, \dots, \mu_k$, a *contrast* among the population means is given by

$$L = c_1\mu_1 + c_2\mu_2 + \dots c_k\mu_k$$

where the sum of the coefficients (c_i) must equal 0. Note, first of all, that all the paired comparisons tested by the Tukey procedure are contrasts. Why? The general form for a paired comparison is $\mu_i - \mu_j$, for the i th and j th groups. The coefficient for μ_i is 1, while the coefficient for μ_j is -1 . But the sum of these coefficients is 0 and therefore we have a contrast.

To illustrate the wide variety of contrasts possible with the Scheffé, consider a 4 group problem, with population means μ_1, μ_2, μ_3 , and μ_4 . First, we can test all paired contrasts for significance (as with the Tukey). Recall that there are 6 paired comparisons (1 vs. 2, 1 vs. 3, 1 vs. 4, 2 vs. 3, 2 vs. 4, and 3 vs. 4). But, in addition, we can test all kinds of complex *contrasts* involving more than two groups for significance. Below we list just four possible complex contrasts:

$$L_1 = \mu_1 - (\mu_2 + \mu_3)/2$$

$$L_2 = (\mu_1 + \mu_2) - (\mu_3 + \mu_4)$$

$$L_3 = \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$$

$$L_4 = \mu_2 - (\mu_3 + \mu_4)$$

Remember that for each of the above to be a contrast the sum of the coefficients must be 0. We show this below for the first two and leave as an exercise for the reader to show that L_3 and L_4 are contrasts.

For L_1 the coefficients are $c_1 = 1$, $c_2 = c_3 = -.5$. Therefore, $c_1 + c_2 + c_3 = 1 + (-.5) + (-.5) = 0$, and L_1 is a contrast. For L_2 the coefficients are $c_1 = c_2 = 1$ and $c_3 = c_4 = -1$. Again, $c_1 + c_2 + c_3 + c_4 = 1 + 1 + (-1) + (-1) = 0$, and L_2 is a contrast.

As with the Tukey procedure, the Scheffé method establishes a set of simultaneous confidence intervals for all population mean contrasts. The lower and upper limits for the intervals are:

$$\hat{L} - \hat{\sigma}_{\hat{L}} \sqrt{(k-1)F_{\alpha; k-1; N-k}} \quad \text{and} \quad \hat{L} + \hat{\sigma}_{\hat{L}} \sqrt{(k-1)F_{\alpha; k-1; N-k}} \quad (12)$$

where \hat{L} is the estimate of the contrast and $\hat{\sigma}_{\hat{L}}$ is the estimated standard error of the contrast. Now, the estimate for a general contrast is obtained by replacing the population means by sample means, that is, $L = c_1\bar{x}_1 + c_2\bar{x}_2 + \dots + c_k\bar{x}_k$. In the Appendix of this chapter we show that the estimated variance for a contrast is given by

$$\hat{\sigma}_{\hat{L}}^2 = MS_w \left(\sum c_i^2 / n_i \right) \quad (13)$$

where MS_w is the denominator of the F test and the n_i represents the number of subjects in the i th group. The standard error of the contrast is simply the square root of Equation 13. To illustrate calculation of a few Scheffé intervals we reconsider the data example used for the Tukey procedure in Section 2.11. There were 4 groups with differing group sizes and $MS_w = 22.22$.

	1	2	3	4
n_i	20	17	14	18
\bar{x}_i	7	8	10	13

We test the following contrasts for significance at an experimentwise error rate = .10, i.e., we will obtain the 90% simultaneous confidence intervals:

$$L_1 = \mu_1 - (\mu_2 + \mu_3)/2 \quad \text{and} \quad L_2 = \mu_2 - (\mu_3 + \mu_4)/3$$

The estimates for contrasts L_1 and L_2 are given by

$$\hat{L}_1 = 7 - (8 + 10)/2 = -2 \quad \text{and} \quad \hat{L}_2 = 7 - (8 + 10 + 13)/3 = -3.33$$

The standard error for \hat{L}_1 is:

$$\hat{\sigma}_{\hat{L}_1} = \sqrt{22.22[(1^2)/20 + (-.5)^2/14]} = 1.355$$

and the standard error for \hat{L}_2 is:

$$\hat{\sigma}_{\hat{L}_2} = \sqrt{22.22[(1^2)/20 + (-.33)^2/17 + (-.33)^2/14 + (-.33)^2/18]} = 1.25$$

$$\text{Also, } \sqrt{(4-1)F_{.10;3,65}} = \sqrt{3(2.18)} = 2.56.$$

The confidence interval for L_1 is given by

$$(-2 - 1.355(2.56), -2 + 1.355(2.56)) \text{ or } (-5.469, 1.469)$$

while the confidence interval for L_2 is given by

$$(-3.33 - 1.25(2.56), -3.33 + 1.25(2.56)) \text{ or } (-6.53, -.13)$$

Recall that if a confidence interval covers 0 it means the contrast is *not* significant. Therefore, L_1 is not significant. However, L_2 is significant since that interval does not cover 0.

2.13 HETEROGENEOUS VARIANCES AND UNEQUAL GROUP SIZES

As previously indicated, the analysis of variance is robust against unequal population variances provided that the group sizes are equal or approximately equal. When heterogeneous variances are present various procedures have been recommended: Welch (1951), Brown and Forsythe (1974) and the Kruskal—Wallis nonparametric test. A Monte Carlo study by Tomarkin and Serlin (1986) examined the above three procedures and found that the Welch test was superior in most cases studied in terms of better control on type I error and greater power.

In terms of post hoc procedures, recall that the Tukey maintained an honest experimentwise error rate with unequal group sizes *only* if the homogeneity of variance assumption is tenable *and* the assumed common n in the Tukey test statistic is replaced by the harmonic mean for each pair of groups. Fortunately, there is also a Welch t statistic which does *not* assume equal variances. Games and Howell (1976), in a Monte Carlo study on the Tukey procedure, found that the Welch approximate t statistic kept the experimentwise error rate under control when heterogeneous variances and unequal group sizes are both present. The Welch approximate t , which we denote by t_w , is given by

$$t_w = (\bar{x}_i - \bar{x}_j) / \sqrt{s_i^2/n_i + s_j^2/n_j}$$

where s_i^2 and s_j^2 are the sample variances for the i th and j th groups and n_i and n_j are the respective group sizes. Note that since the homogeneity of variance of assumption is not tenable the Welch statistic uses only those variances for the pair of groups being compared. A pooled error term would be inappropriate since the

sample variances are estimating different population values. The degrees of freedom (v) for each Welch statistic will in general be different and is given by

$$v = \frac{(s_i^2 / n_i + s_j^2 / n_j)^2}{\frac{(s_i^2 / n_i)^2}{n_i - 1} + \frac{(s_j^2 / n_j)^2}{n_j - 1}}$$

A pair of means was declared to be significantly different in the Games and Howell study if

$$|t_w| > q_{\alpha,k,v} / \sqrt{2}$$

Note that with the above approach several different critical values from the studentized range table will be needed, since v will tend to be different for the various paired comparisons.

Below are a few selected results from their study comparing the Welch statistic against a pooled error term (MS_w) approach for a 4 group situation:

<i>Group Sizes</i>	<i>MS_w</i>	<i>Welch</i>
16, 14, 10, 6		
<i>Population Variances</i>		
1, 3, 5, 7	.122	.060
1, 1, 7, 7	.115	.064
1, 1, 1, 13	.112	.064

These error rates are to be compared against a significance level of .05. The above situations all represent positively biased situations, i.e., where the large variances are associated with the small group sizes. Recall that for ANOVA these are the situations where it was liberal, with the error rate greater than level of significance. The above results showed that use of a pooled error term caused the Tukey approach to be quite liberal, i.e., the actual error rate was over twice the significance level, while use of the Welch statistic kept the actual α quite close to the significance level of .05.

Example

To illustrate with an example, consider the four group data set in the table on the next page.

These data were run on SPSS for Windows 12.0, and a selected printout from that run is given in Table 2.7.

Group 1	Group 2	Group 3	Group 4
14	20	36	26
21	25	29	35
37	18	31	46
18	30	22	18
20	26	45	30
29	22	43	33
42	31	27	49
12	26	33	15
27	28	35	27
30	24	28	
33	19	36	
	17	40	
	21	38	
	23	29	
	27	22	
	32		
	19		
	29		
	28		
	23		

2.14 MEASURES OF ASSOCIATION (VARIANCE ACCOUNTED FOR)

One of the facts of “statistical life” is that whether we obtain significance with *any* statistical test is *heavily* dependent on sample size. With large enough sample sizes even very small differences among the group means will be declared statistically significant. Why is this so? To shed some light on this it is helpful to first recall from Section 2.5 that the numerator of the F statistic, for equal group sizes, can be written as

$$MS_b = n \sum (\bar{x}_i - \bar{x})^2 / (k - 1)$$

Thus, the F ratio can be written as

$$F = \frac{n \sum (\bar{x}_i - \bar{x})^2 / (k - 1)}{MS_w}$$

Assuming the null hypothesis is false, the numerator can be made arbitrarily large by increasing the group sample size n . Now, increasing the sample sizes should not have any systematic effect on MS_w , and we assume for the sake of sim-

plicity that MS_w remains the same. But, given the above two statements, we see that F can be made arbitrarily large by increasing sample size. We now consider a numerical example to illustrate the above. Suppose two studies are done, each with 3 groups, and the group means in both cases are 10, 14, and 18. Assume $MS_w = 100$ in both cases. One study has 16 subjects per group while the other has 100 subjects per group. The grand mean = 14, and the F for the first study is

$$F = 16/2 [(10 - 14)^2 + (18 - 14)^2]/100 = 2.56$$

The critical value for significance at .05 is 3.15, and therefore this result would not be significant.

The other study, *with the same mean differences*, has an F of

$$F = 100/2[(10 - 14)^2 + (18 - 14)^2]/100 = 16$$

and this would be significant well beyond the .001 level!

Now, we illustrate the other point. That is, even very small differences will be declared significant if n is large enough. Suppose again 3 groups with $MS_w = 100$, but now there are 400 subjects per group. The means for the groups are 10, 11 and 12. The F ratio here will be $F = 200[(10-11)^2 + (12-11)^2]/100 = 4$, which is significant at the .05 level, even though the mean differences are very small. To use a domestic analogy, this is like using a sledgehammer to pound out significance.

Because of this kind of situation it has been argued for some time (perhaps popularized most by Hays in his influential text *Statistics for Psychologists*, 1963) that we need some way of determining whether a statistically significant result is practically significant. Hays (1963) introduced his $\hat{\omega}^2$ as a measure of association (strength of relationship) to get at practical significance, and such measures have subsequently been recommended by various textbook authors (Cohen & Cohen, 1975; Kerlinger & Pedhazur, 1973; Kirk, 1982). The two most commonly used measures of this type are η^2 and Hays $\hat{\omega}^2$. The formulas for each of them for a one way ANOVA are given below:

$$\eta^2 = SS_b / SS_t$$

where $SS_t = SS_b + SS_w$ (total sum of squares), and

$$\hat{\omega}^2 = (SS_b - (k - 1)MS_w) / (SS_b + MS_w)$$

Usually the numerical values for these two will not differ a great deal, although Hays' measure is generally regarded as preferable because he used unbiased estimates in deriving his measure (but the measure itself is *not* unbiased).

Such measures can be useful after the test of significance, since they are essentially independent of sample size. Nevertheless, there are limitations associated with these measures, as O'Grady (1982) has pointed out in an excellent review on

measures of explained variance. He cites 3 basic reasons why such measures should be interpreted with caution (measurement, methodological, and theoretical). With respect to measurement he notes that the reliability of the variables somewhat restricts how large a measure of association can be, since these measures are correlational in nature. Several methodological factors are mentioned; we discuss just two of them. One is the homogeneity of the population sampled. Since measures of association are correlational measures, the more homogeneous the population, the smaller the correlation will tend to be, and therefore the smaller the percent of variance that can be potentially accounted for. This is simply the *restriction of range* phenomenon you encountered in beginning statistics when studying the Pearson correlation. A second factor that can have a substantial effect on the magnitude of a measure of association is the number of levels chosen, and how they are chosen, for a fixed effects ANOVA (which is what we are dealing with). To illustrate he uses the following example. Suppose there are 3 researchers that wish to examine the relationship between a hypothesized carcinogen and the incidence of cancer. The first researcher chooses to contrast a control condition (0 exposure) with a 2% exposure to the carcinogen. The second researcher chooses to maximize

TABLE 2.7
Selected Printout From SPSS for Windows 12.0 for Unequal Variances
Example With Tamhane and Games–Howell Post Hoc Procedures

Insert art from p. 94 of previous edition

TABLE 2.7
(Continued)

Insert art from p. 95 of previous edition

the changes of a relationship and contrasts 0% exposure with 20% exposure. Finally, the third researcher is interested in determining the shape of the relationship across various levels of the supposed carcinogen. Below we present the descriptive statistics, *F* ratios, and eta squares for the 3 studies:

		<i>Exposure Group</i>					
<i>Study</i>		0%	2%	5%	10%	20%	
1	\bar{x}	10	12				$F = 5, \eta^2 = .22$
	<i>s</i>	2	2				
2	\bar{x}	10				18	$F = 80, \eta^2 = .82$
	<i>s</i>	2				2	
3	\bar{x}	10	12	14	16	18	$F = 25, \eta^2 = .69$
	<i>s</i>	2	2	2	2	2	

Ten subjects are assumed in each group in each study. Notice how the measure of association is drastically affected by the number of levels and how the levels are chosen, even though the means and standard deviations are the same.

A theoretical point O'Grady mentions which should be kept in mind before casting asperations on a "low" amount of variance accounted for is that most behaviors have *multiple causes*, and hence it will be difficult in these cases to account for a large amount of variance with just a single cause (say treatments).

Anyone planning on using measures of association in their research should read and think carefully about O'Grady's paper. To enforce the point that a "small" amount of variance accounted for may indeed be practically significant we consider an example from Rosenthal and Rosnow (1984). They consider the comparison of a treatment and control group where the dependent variable is dichotomous, whether the subjects live or die. The following table is presented:

	<i>Treatment Outcome</i>		
	Alive	Dead	
Treatment	66	34	100
Control	34	66	100
	100	100	

Since both variables are dichotomous, the phi coefficient ϕ , a special case of the Pearson correlation for dichotomous variables (Glass & Hopkins, 1984), measures the relationship between them:

$$\phi = (34^2 - 66^2) / \sqrt{100(100)(100)(100)} = -.32$$

Squaring ϕ (since it is a correlation) yields variance accounted for, which is $(-.32)^2 = .10$. Thus, the treatment-control distinction accounts for "only" 10% of the variance in the outcome. However, this is enough to increase the survival rate from 34% to 66%, far from trivial! The same type of interpretation would hold if we were to consider some less dramatic type of outcome like improvement vs. no improvement, where treatment was, say, a type of psychotherapy. Also, the interpretation is *not* confined to just a dichotomous measure.

2.15 PLANNED COMPARISONS

One approach to the analysis of data is to first demonstrate overall significance, and then follow up to assess the significant subsources of variation (i.e., which particular groups differed). This approach is appropriate in *exploratory* studies where it is necessary to first establish that an effect exists. There may be a weak literature

base, or none on which to base specific hypotheses. This type of study is somewhat unfocused and some have even referred to these studies as “fishing expeditions.”

Now we consider a more focused type of study, where there either is a fairly strong theoretical and/or literature base, or the investigator has specific questions to ask of the data. These questions will be in the form of hypotheses involving group comparisons. This is more of a *confirmatory* type study. Here, a priori, the investigator sets up planned comparisons among the group means. It is important to use planned comparisons when the situation justifies them, since performing a small number of statistical tests cuts down on the probability of spurious results (type I errors).

Hays (1981) has shown that planned comparisons are a more powerful approach statistically. If we set up a small number of comparisons, then power will be enhanced and overall α can be controlled through the *Bonferroni Inequality*. This is a very important inequality. It states that if k hypotheses, k planned comparisons here, are tested separately with type I error rates of $\alpha_1, \alpha_2, \dots, \alpha_k$, then

$$\text{overall } \alpha \leq \alpha_1 + \alpha_2 + \dots + \alpha_k \quad (16)$$

If the hypotheses are each tested at the same alpha level, say α' , then the Bonferroni upper bound becomes

$$\text{overall } \alpha \leq k\alpha' \quad (17)$$

If the comparisons are independent (this is defined shortly), then an *exact* calculation for overall α is available. First, $(1 - \alpha_1)$ is the probability of no type I error for the first comparison. Similarly, $(1 - \alpha_2)$ is the probability of no type I error for the second, $(1 - \alpha_3)$ the probability of no type I error for the third, etc. If the tests are independent, then we can multiply probabilities. Therefore, $(1 - \alpha_1)(1 - \alpha_2)\dots(1 - \alpha_k)$ is the probability of *no* type I errors for all k tests. Thus,

$$\text{overall } \alpha = 1 - (1 - \alpha_1)(1 - \alpha_2)\dots(1 - \alpha_k) \quad (18)$$

is the probability of at least one type I error. If the tests are not independent, then overall α will still be *less* than given in Equation 18 although it is very difficult to calculate. If we set the alpha levels equal, say to α' , for each test, then Equation 18 becomes overall $\alpha = 1 - (1 - \alpha')(1 - \alpha')\dots(1 - \alpha') = 1 - (1 - \alpha')^k$. This expression, $1 - (1 - \alpha')^k$, is approximately equal to $k\alpha'$ for small α' . The table below compares the two for $\alpha' = .05, .01$, and $.001$ for a number of tests ranging from 5 to 100.

<i>No. of Tests</i>	$\alpha' = .05$		$\alpha' = .01$		$\alpha' = .001$	
	$1 - (1 - \alpha')^k$	$k\alpha'$	$1 - (1 - \alpha')^k$	$k\alpha'$	$1 - (1 - \alpha')^k$	$k\alpha'$
5	.226	.25	.049	.05	.00499	.005
10	.401	.50	.096	.10	.00990	.010
15	.537	.75	.140	.15	.0149	.015
30	.785	1.5	.260	.30	.0296	.030
50	.923	2.5	.395	.50	.0488	.050
100	.994	5.0	.634	1	.0952	.100

First, the numbers in the table greater than 1 don't represent probabilities, since a probability can't be greater than 1. Second, note that if we are testing each of a large number of hypotheses at the .001 level, the difference between $1 - (1 - \alpha')^k$ and the Bonferroni upper bound of $k\alpha'$ is very small and of no practical consequence. Also, the differences between $1 - (1 - \alpha')^k$ and $k\alpha'$ when testing at $\alpha' = .01$ are also small for up to about 30 tests. For more than about 30 tests $1 - (1 - \alpha')^k$ provides a tighter bound and should be used. When testing at the $\alpha' = .05$ level, $k\alpha'$ is okay up for about 10 tests, but beyond that $1 - (1 - \alpha')^k$ is much tighter and should be used.

Example 1

We have a 4 group problem with 3 planned comparisons and want overall $\alpha < .10$. This can be achieved by simply dividing the overall α by the number of tests done, i.e., $.10/3 = .033$. Thus, if each comparison is tested at the $\alpha = .033$ level, we are assured by the Bonferroni inequality that

$$\text{overall } \alpha \leq .033 + .033 + .033 = .10$$

Example 2

Suppose there are 5 planned comparisons in a 6 group problem. If we test the first two at the .05 level (i.e., $\alpha_1 = .05$ and $\alpha_2 = .05$), and the remaining three comparisons at the .01 level ($\alpha_3 = \alpha_4 = \alpha_5 = .01$), then we are assured by the inequality that

$$\text{overall } \alpha \leq .05 + .05 + .01 + .01 + .01 = .13$$

Now let us consider a couple of research examples of setting up planned comparisons. The next sample has treatments of a structure that could be useful in a variety of fields.

Example 3

Consider a four group situation involving a comparison of two treatments, a combination of the two treatments, and a control group on some dependent measure. Schematically, we have

T_1 (control)	T_2	T_3	T_4 (T_2 and T_3 combined)
μ_1	μ_2	μ_3	μ_4

The two treatments might be two reading methods, two types of counseling, two diets, etc. Of course, the two treatments would have to be such that combining them made sense. Now there are three very meaningful, focused questions to ask of the data.

1. Is something better than nothing? Here we are comparing the control group vs. the treatment groups.
2. Do the two individual treatments differ in effectiveness?
3. Is the combination of treatments more effective than either treatment individually?

These comparisons will be set up as *contrasts* among the population means for the groups. In general, for k groups with population means $\mu_1, \mu_2, \dots, \mu_k$, a contrast among the population means is given by

$$L = c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k$$

where the sum of the coefficients (c_i) must equal 0.

Let us set up the comparisons for the 3 questions above, and we will see that each is a contrast:

$$L_1 = \mu_1 - (\mu_2 + \mu_3 + \mu_4)/3$$

The coefficients here are $c_1 = 1$, $c_2 = c_3 = c_4 = -1/3$. The sum of these coefficients = 0, so L_1 is a contrast.

$$L_2 = \mu_2 - \mu_3$$

The coefficients are $c_2 = 1$ and $c_3 = -1$, so that $c_2 + c_3 = 0$ and L_2 is a contrast.

$$L_3 = \mu_4 - (\mu_2 + \mu_3)/2$$

Here $c_4 = 1$, $c_2 = c_3 = -.5$, so that $\sum c_i = 0$, and L_3 is a contrast.

The formula for the sum of squares of a contrast is given by

$$SS_L = \hat{L}^2 / \sum c_i^2 / n_i$$

Where \hat{L} is the estimate of the contrast and the n_i are the group sizes.

For *equal* group size the above set of three contrasts represent *orthogonal* comparisons. The sums of squares associated with the contrasts (denote them by SS_L) are independent, and it can be shown that:

$$SS_b = SS_{L_1} + SS_{L_2} + SS_{L_3} \quad (19)$$

That is, the overall between groups variation is additively partitioned into three independent pieces of variation. For equal group size the condition that needs to be met for a pair of contrasts to be independent is that the sum of the products of the coefficients equal 0. We now show that this condition is met for all three pairs of contrasts. We present the contrasts below in schematic form, just using the coefficients that define each contrast:

	T_1	T_2	T_3	T_4
L_1	1	-1/3	-1/3	-1/3
L_2	0	1	-1	0
L_3	0	-1/2	-1/2	1

The sum of the products of the coefficients for each pair are:

$$L_1 \text{ and } L_2: 1(0) + (-1/3)(1) + (-1/3)(-1) + (-1/3)(0) = 0$$

$$L_1 \text{ and } L_3: 1(0) + (-1/3)(-1/2) + (-1/3)(-1/2) + (-1/3)(1) = 0$$

$$L_2 \text{ and } L_3: 0(0) + 1(-1/2) + (-1)(-1/2) + 0(1) = 0$$

There are other sets of three orthogonal comparisons for this problem, and *any* such set will also provide for an additive partitioning of the SS_b . The nice feature about orthogonal contrasts is that significance on one contrast implies nothing about potential significance on another contrast. That is, we do *not* have a confounding of the sources of variation. With correlated contrasts the sources of variation are confounded; however, the unique sum of squares associated with each contrast can be obtained by using the SPSS MANOVA program, which since Release 2.2 has the unique sum of squares as the default option. *Although it is desirable to have orthogonal contrasts, the set of contrasts to impose in a given situation should be dictated by the research questions of the investigator.*

Now we express the condition for independence of two contrasts for equal group size in general form. Consider two general contrasts for k groups:

$$L_1 = c_{11}\mu_1 + c_{12}\mu_2 + \dots + c_{1k}\mu_k$$

$$L_2 = c_{21}\mu_1 + c_{22}\mu_2 + \dots + c_{2k}\mu_k$$

The condition for independence is

$$c_{11}c_{21} + c_{12}c_{22} + \dots + c_{1k}c_{2k} = 0$$

If the group sizes are not equal, then the condition for independence is more complicated and becomes:

$$(c_{11}c_{21})/n_1 + (c_{12}c_{22})/n_2 + \dots + (c_{1k}c_{2k})/n_k = 0$$

Example 4

A medical researcher wishes to evaluate the effectiveness of 4 drugs on reaction time. Schematically, the design is

Control	One Generic Type Drug A	Drug B	Other Generic Type Drug C	Drug D
μ_1	μ_2	μ_3	μ_4	μ_5

A set of 4 focused and relevant questions to ask here are:

1. Are drugs more effective than no drugs?

$$L_1 = \mu_1 - (\mu_2 + \mu_3 + \mu_4 + \mu_5)/4$$

2. Is one generic type of drug more effective than the other generic type?

$$L_2 = (\mu_2 + \mu_3)/2 - (\mu_4 + \mu_5)/2$$

3. Are the two drugs of the first generic type different in effectiveness?

$$L_3 = \mu_2 - \mu_3$$

4. Are the two drugs of the other generic type different in effectiveness?

$$L_4 = \mu_4 - \mu_5$$

The reader should verify that each of these comparisons is indeed a contrast.

2.16 TEST STATISTIC FOR PLANNED COMPARISONS

Recall that for k groups with population means $\mu_1, \mu_2, \dots, \mu_k$, a contrast (L) among the population means is given by

$$L = c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k$$

where $\sum c_i = 0$.

This contrast is estimated by replacing the population means by the sample means, yielding

$$\hat{L} = c_1\bar{x}_1 + c_2\bar{x}_2 + \cdots + c_k\bar{x}_k$$

To test whether a given contrast is significantly different from 0, i.e., to test

$$H_0: L = 0 \text{ vs. } H_1: L \neq 0$$

we need an expression for the variance of a contrast. We show in the Appendix at the end of this chapter that the variance for a contrast is given by

$$\hat{\sigma}_{\hat{L}}^2 = MS_w(\sum c_i^2 / n_i)$$

where MS_w is the error term from all the groups (the denominator of the F test) and the n_i are the group sizes.

Therefore, the following F statistic can be used to test a contrast for significance,

$$F = \hat{L}^2 / (MS_w(\sum c_i^2 / n_i)) = \frac{\hat{L}^2 / \sum c_i^2 / n_i}{MS_w} \quad (20)$$

with 1 and $(N-k)$ degrees of freedom. Note here that each contrast has just *one* degree of freedom. As Hays and others have indicated, the $(k-1)$ between group degrees of freedom can be partitioned into $(k-1)$ non-redundant single degree of freedom contrasts.

Note that if the group sizes are equal ($n_1 = n_2 = \cdots = n_k = n$), then Equation 20 can be written in somewhat simpler form as:

$$F = \frac{n\hat{L}^2 / \sum c_i^2}{MS_w}$$

Also, some authors present the test statistic for planned comparisons as a t statistic:

$$t = \frac{\hat{L}}{\sqrt{MS_w \sum c_i^2 / n_i}} \quad (21)$$

Since it can be shown that $F = t^2$, the F test is equivalent to a two tailed t test at the same level of significance. SPSS, as shown in Table 2.10, uses the t statistic for

contrasts. The probabilities given are for a *two tailed* test, so that if you are testing a directional hypothesis with your contrast the probability value should be divided by two.

Numerical Example

Suppose an investigator has a 4 group problem and wishes to examine the following planned comparisons:

	Groups				
	1	2	3	4	(in population means)
L_1	1	-.5	-.5	0	$L_1 = \mu_1(\mu_2 + \mu_3)/2$
L_2	.5	.5	-.5	-.5	$L_2 = (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$
L_3	0	0	1	-1	$L_3 = \mu_3 - \mu_4$

Notice that on the left the contrasts are indicated schematically by simply using the coefficients for the population means. This is the way contrasts are input for SPSS, and SAS.

Suppose we have the following descriptive information for the groups:

	1	2	3	4
n_i	10	8	11	13
\bar{x}_i	5.6	7.3	8.1	4.2

and it is known that the pooled error term for the groups is $MS_w = 8.7$.

We now show how to calculate the test statistic given in Equation 20 for testing contrasts 1 and 3, and leave the calculation of contrast 2 as an exercise for the reader.

Contrast 1

First we obtain the estimate of the contrast using the sample means:

$$\hat{L}_1 = 5.6 - (7.3 + 8.1)/2 = 5.6 - 7.7 = -2.1$$

Also, we have

$$\begin{aligned} \sum c_i^2 / n_i &= 1^2 / 10 + (-.5)^2 / 8 + (-.5)^2 / 11 = .154 \\ F &= \frac{(-2.1)^2 / .154}{8.7} = 3.29 \end{aligned}$$

The critical value at .05 for this contrast is $F_{.05;1,38} = 4.08$. Therefore, this contrast would not be significant. Note that from the above critical value the between degrees of freedom for the contrast is 1.

Contrast 3

The estimate of the contrast is given by

$$\hat{L}_3 = 8.1 - 4.2 = 3.9$$

Also, we have

$$\sum c_i^2 / n_i = 1^2 / 11 + (-1)^2 / 13 = .168$$

Thus,

$$F = \frac{(3.9)^2 / .168}{8.7} = 10.406$$

The critical value for this contrast at the .05 level is the *same* as for contrast 1, i.e., 4.08. Thus, contrast 3 is significant.

2.17 PLANNED COMPARISONS ON SPSS AND SAS

To illustrate how to set up planned comparisons on the statistical packages, and how to interpret the output, we ran the data for Example 4 (Section 2.15) involving the effect of different generic type drugs on reaction time. The complete SAS control lines are given in Table 2.8. To obtain the planned comparisons on SPSS we used the Windows 12.0 version.

To obtain planned comparisons on SPSS with windows is a simple matter. One first goes to and clicks on STATISTICS, then on COMPARE MEANS, and finally click on ONE WAY ANOVA. When all of this is done the first screen displayed in Table 2.9 appears. Make REACTIME the dependent variable and DRUG the factor (grouping variable), and then click on CONTRASTS. When this is done the middle screen in Table 2.9 appears. Recall that for this example we have 5 groups, so we can have at most 4 contrasts. enter a coefficient for each group (category) of the factor variable and click **ADD** after each entry. Each new value is added to the bottom of the coefficient list. To specify additional contrasts, click **NEXT**. For the current example, when the last contrast is entered the bottom screen appears. Click on CONTINUE and then click on OK to run the contrasts.

Selected annotated printout from the SPSS and SAS runs is presented in Table 2.10.

TABLE 2.8
SAS Control Lines for Planned Comparisons on Drug Data*

```
DATA CONTRAST;  
INPUT DRUG REACTIME @@;  
LINES;  
1 5 1 8 1 8 1 11 1 1 1 9  
1 5 1 9  
2 16 2 18 2 5 2 12 2 11 2 12  
2 23 2 19  
3 16 3 7 3 10 3 4 3 7 3 23  
3 12 3 13  
4 2 4 10 4 9 4 13 4 11 4 9  
4 13 4 9  
5 7 5 11 5 12 5 9 5 14 5 16  
5 24 5 19  
PROC PRINT;  
PROC MEANS;  
BY DRUG;  
PROC GLM;  
CLASS DRUG;  
MODEL REACTIME = DRUG;  
CONTRAST 'DRUG VS NO DRUG' DRUG 1 -.25 -.25 -.25 -.25;  
CONTRAST 'GENTYPE1 VS GENTYPE2' DRUG 0 .5 .5 -.5 -.5;  
CONTRAST 'DRUG A VS DRUG B' DRUG 0 1 -1 0 0;  
CONTRAST 'DRUG C VS DRUG D' DRUG 0 0 0 1 -1;
```

* Recall that the design was

One Generic Type			Other Generic Type	
Control	Drug A	Drug B	Drug C	Drug D
μ_1	μ_2	μ_3	μ_4	μ_5

2.18 THE EFFECT OF AN OUTLIER ON AN ANOVA

In Chapter 1 we indicated the importance of outliers and showed that they can have a dramatic effect on the results of a statistical analysis. Here we illustrate that effect for a one way analysis of variance.

Example

An investigator has collected the following data:

<i>Gp 1</i>	<i>Gp 2</i>	<i>Gp 3</i>
15	17	6
18	22	9
12	15	12
12	12	11
9	20	11
10	14	8
12	15	13
20	20	30
	21	7

The score of 30 in group 3 is an outlier. With that case in the ANOVA we do not find significance ($F = 2.61, p < .095$) at the .05 level, while with the case *deleted* we do find significance well beyond the .01 level ($F = 11.18, p < .0004$). Deleting the case has the effect of producing greater separation among the means, since the means with the case included are (13.5, 17.33, 11.89), while the means with the case deleted are (13.5, 17.33, 9.63). It also has the effect of reducing the within group variability in group 3 substantially, and hence the pooled within group variability (the error term for ANOVA) will be much smaller.

2.19 MULTIVARIATE ANALYSIS OF VARIANCE

In this chapter we have considered what is called *univariate* analysis of variance, since there is just one dependent variable in the analysis. In many studies, however, the subjects are measured on several variables. The appropriate statistical analysis for comparing the k groups on the p dependent variables *simultaneously* is called multivariate analysis of variance (MANOVA). This type of analysis is to be distinguished from doing a separate univariate ANOVA on each dependent variable. Four reasons why a MANOVA is preferable to such separate univariate analyses are:

1. The univariate analyses, especially for a moderate or large number of dependent variables, allow the overall type I error rate to go completely out of control. The situation here is analogous to what happened when doing several t tests for the k group problem.
2. The univariate ANOVAs ignore important information, such as the correlations among the dependent measures, whereas the multivariate tests incorporate these correlations into the test statistics.
3. The univariate tests may not show the groups to be significantly different on any of the variables, because of small unreliable differences on each of the variables. However, if measures are considered *jointly* (as in MANOVA) there may be

TABLE 2.9
SPSS for Windows 12.0 Screens for Obtaining Planned Comparisons

Insert art from p. 108 of previous edition

It can be shown that $t^2 = F$, and note here that $(-2.766)^2 = 7.65$, $(.719)^2 = .52$, etc.

②Using $\alpha = .0125$, we find by examining the tail probabilities that only contrast 1 is significant.

91

a significant difference. That is, small differences on each of the variables may combine to produce a reliable overall difference.

4. If treatments affect the dependent variables in different ways, and the dependent variables are at least moderately correlated within groups, the multivariate approach will be quite powerful and can detect differences that the univariate tests cannot. One of the exercises illustrates this situation.

2.20 SUMMARY

1. The analysis of variance (ANOVA) is appropriate for comparing k independent groups on a single dependent variable. It is the generalization of the t test, which is used to compare two groups.

2. It was shown how the use of multiple t tests for the k group problem allows the overall α level to get out of control, hence the need for ANOVA.

3. In testing the null hypothesis of equal population means, the ANOVA computes and compares two basic sources of variation (between and within). Between group variability measures variability of the group mean about the grand mean, while within group variability measures how much subjects vary who are treated alike.

4. It was shown that MS_w and MS_b both represent variances.

5. Analysis of variance rests on three assumptions: normality of scores in each group, equal population variances, and independence of the observations. Considerable research on violations of assumptions suggests that ANOVA is robust with respect to a violation of the normality assumption. It is robust against unequal variances provided that the group sizes are equal or approximately equal (largest/smallest < 1.5). ANOVA is severely affected by correlated observations. Two methods are suggested for dealing with correlated observations. One is to simply test at a more stringent α level. The other, if dealing with several small groups within each treatment, is to use the group mean as the unit of analysis.

6. After a significant F , several post hoc procedures are mentioned for locating where the differences lie. For most situations the Tukey procedure is preferred because of 3 reasons: (a) it does control the overall α , (b) it is fairly powerful for detecting differences, and (c) it examines a meaningful and easily interpreted set of comparisons (pairwise comparisons). For more extensive data-snooping, the Scheffé procedure should be used. This procedure is quite conservative, however, so for more adequate power one will either need to have a large number of subjects or set overall α at .10 or .15. The Dunnett procedure should be used if you are only interested in comparing each treatment group against the control group.

7. For situations where the homogeneity of variance assumption is not tenable and there are sharply unequal group sizes, an ANOVA and post hoc procedure that

do not assume equal variances should be used. In Section 2.13 we illustrated two post hoc procedures (Tamhane and Games-Howell), which do not assume equal variances.

8. Confidence intervals and measures of associations (variance accounted for) are mentioned as two ways of determining the practical significance of a study. Several cautions to be observed in using measures of association are mentioned, and an example is given to illustrate that a “small” amount of variance accounted for could indeed be practically significant.

9. Planned comparisons are presented as an alternative way to analyze the k group problem. Here the researcher a priori is setting up comparisons among the group means corresponding to his or her hypotheses. An overall F test is *not* required in this approach. Planned comparisons are a more powerful approach, and overall α can be controlled through use of the Bonferroni Inequality.

EXERCISES

- For a 7 group problem there would be 21 t tests for the 21 paired comparisons. Using this approach, rather than the proper one way ANOVA, what would overall α be approximately (if each t test is done at .05 level)?
- If $k = 4$ and $n_1 = n_2 = n_3 = n_4 = 20$, then $df_h = ?$, $df_w = ?$
 - If $k = 3$ and $n_1 = 12$, $n_2 = 25$, and $n_3 = 20$, then $df_h = ?$ $df_w = ?$
- Find the critical value at the .05 level for a 3 group problem with 10 subjects per group.
 - Find the critical value at the .10 level for a 4 group problem with 20 subjects per group.
 - Find the critical value at the .01 level for a 5 group problem with 8 subjects per group.
- Do a one way analysis of variance on the following data, testing for significance at the .10 level.

Treat 1	Treat 2	Treat 3
11	10	15
14	8	14
13	12	10
17	7	11
18	13	16

(b) Obtain MS_w by computing the variances for each group using your TI-30Xa STAT calculator, or some other calculator that yields means and variances.

5. Do a one way ANOVA on this data; test for significance at .05:

Group 1	Group 2	Group 3	Group 4
2	7	4	8
3	9	4	4
5	11	5	7
6		8	7
		3	

In obtaining MS_w with your calculator, note that

$$SS_w = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2$$

where $s_1^2, s_2^2, \dots, s_k^2$ are sample variances for groups.

6. As part of a study by Sarachen-Deily (1985), deaf high school students were classified as good readers, average readers or poor readers on the basis of scores on the Stanford Achievement Test, Special Edition for Hearing Impaired Students. The following table resulted:

Category	<i>n</i>	Mean	Stand. Deviation
Good reader	7	6.90	.49
Aver. reader	9	5.04	.56
Poor reader	4	3.38	.10

An ANOVA on this data yielded a significant overall difference at the .01 level.

- (a) Are the sample variances for the groups sharply unequal?
- (b) Would you be worried about the significant ANOVA result being spurious? Why, or why not?

7. A 4-group ANOVA is run on the following data:

	Gp1	Gp2	Gp3	Gp4
n_i	15	15	15	15
\bar{x}_i	5.6	7.3	4.1	8.7

The F statistic is $F = 63.73/22.35 = 2.85, p < .10$.

Apply the Tukey procedure at the .10 level to determine which pairs of groups are significantly different.

8. A study by Smith, Jones, and Waugh (1986) evaluated the effect of interactive computer assisted videodisc laboratory simulations in enhancing achievement in a freshmen college chemistry course. A group of 103 students were randomly assigned to one of three groups: (a) the first group was required to complete a series of interactive videodisc lessons on chemical equilibrium in place of laboratory work (the VDISC group), (b) the second group was required to complete only a traditional laboratory experiment on the same content material (the LAB group) and (3) the third group was required to complete the interactive videodisc lessons *before* the traditional laboratory experiment (VDISC+LAB group). Following these treatments all students took a seven item multiple choice quiz that required them to apply knowledge of chemical equilibrium to solve both familiar and unfamiliar systems, with the following results:

	VDISC	VDISC+LAB	LAB
n	21	17	49
\bar{x}	5.810	5.588	4.163
sd	1.167	1.121	1.519

A one way ANOVA on these data yielded $F = 13.84, p < .00005$.

- (a) The group sizes are sharply unequal. Would you be worried about the homogeneity of variance assumption? Why, or why not?
- (b) Apply the Tukey procedure with overall $\alpha = .05$ to determine which pairs of groups differ significantly.
- (c) Does there appear to be practical significance? In answering this, calculate the effect size(s) for the pair(s) that are significantly different.
9. One of the planned comparisons involved in the numerical example in 2.16 was

$$L_2 = (\mu_1 + \mu_2)/2 - (\mu_3 + \mu_4)/2$$

Test this contrast for significance at the .05 level.

10. For example 3 in Section 2.15 the following three contrasts were defined: $L_1 = \mu_1 - (\mu_2 + (\mu_3 + \mu_4)/3)$, $L_2 = \mu_2 - \mu_3$, and $L_3 = \mu_4 - (\mu_2 + (\mu_3)/2)$. Label the treatment variable TREAT and the dependent variable DEP. Show the complete control lines for running the above three contrasts on SPSS and on SAS. For the data lines element just put DATA.

11. O'Grady in his paper on measures of association, states

For instance, examining the relationship between the five different types of fitness programs and subsequent efficiency of the heart would, I suspect, produce quite different measures of explained variance for a population of runners in comparison with a population of sedentary individuals (which, in turn, would no doubt produce a quite different value than would be found for a random sample of the American adult population). (p. 773)

What does this relate to that was discussed in the chapter on measures of association?

12. Consider the following data for three groups of subjects on two dependent variables y_1 and y_2 :

Group 1		Group 2		Group 3	
y_1	y_2	y_1	y_2	y_1	y_2
3	7	4	5	5	5
4	7	4	6	6	5
5	8	5	7	6	6
5	9	6	7	7	7
6	10	6	8	7	8

(a) Run a one way ANOVA for y_1 and for y_2 on SAS and a one way multivariate analysis of variance on SAS. All three analyses are obtained in one run. Simply put this in the MODEL statement:

MODEL Y1 Y2 = GPID;

(b) Is there a significant difference for y_1 , at the .05 level? for y_2 at the .05 level?

(c) Are the multivariate tests significant at the .05 level?

(d) In discussing the results from (b) and (c), first look at the pattern of means for y_1 , and y_2 over the three groups. Are the patterns different? Another factor that is important is the within group correlation for y_1 and y_2 , as this has a strong influence on the magnitude of the error term for the multivariate tests (see Stevens, 1986, Chapter 4).

13. An auditor from the Internal Revenue Service wishes to compare the efficiency of four regional tax processing centers. A random sample of 10 returns is selected from each center, and the number of days between receipt

of the tax returns and final processing is determined. The results (in days) are as follows:

East	Midwest	South	West
49	47	39	52
54	56	55	42
40	40	48	57
60	51	43	46
43	55	50	50
65	36	63	34
59	38	48	40
70	52	57	51
61	41	49	39
48	43	65	36

- (a) Run this data on SAS to determine whether there is difference in average processing time among the 4 centers at the .05 level.
 - (b) Are there any significant pairwise differences at the .05 level with the Scheffe procedure? With the Tukey procedure?
 - (c) Explain the difference between the results found in (b) for the two parts of the problem.
14. An investigator randomly assigns 20 subjects to each of four groups (two control and two treatment) and is interested in the effect of treatments on sociability. She wishes to determine whether treatments differ from no treatment, whether the treatment groups do better than the Hawthorne control group, and whether there is a difference in the efficacy of the treatments. Schematically then we have the following contrasts:

	Control	Hawthorne Control	Treat 1	Treat 2
L_1	1	1	-1	-1
L_2	0	1	-.5	-.5
L_3	0	0	1	-1

Is this a set of orthogonal contrasts?

15. Levin, McCormick, Miller, Berry, and Presley (1982) had fourth grade students learn a list of relatively complex English vocabulary words in two experiments. In Experiment 1, pupils used either a mnemonic (keyword) contextual or a verbal contextual procedure. In Experiment 2, three other conditions were compared to the keyword context condition. In Experi-

ment 2 the 64 fourth graders were randomly assigned, 16 to each group, with the following summary statistics resulting:

	Keyword Context	Experiential Context	Picture Context	Control
Mean	72.3	36.2	42.4	48.7
Stand. Dev.	22.9	27.0	23.1	25.6

Levin et al. state in their *Results* section,

Performance differences among conditions were assessed in terms of five planned non-orthogonal comparisons, each based on $\alpha = .01$. Statistical analysis revealed that students in the keyword context condition substantially outperformed those in the control condition, $t = 2.71$, $p < .01$, and the picture context condition, $t = 3.42$, $p < .005$, as well as those in the experiential context condition, $t = 4.14$, $p < .001$. Neither picture context nor experiential context differed significantly from controls, (p. 130)

- (a) What are the 5 planned comparisons?
- (b) Show that they are non-orthogonal.
- (c) Show how the three t values indicated above are obtained.

16. As mentioned in Section 2.10, if your interest in a study is confined to testing each of several treatment groups against a control group, then the Dunnett procedure is the most powerful. Let $t(d)$ represent the modified t value from Dunnett's table (Appendix B.3), n the number of subjects in each group, and MS_w the error term. Then the critical value that must be exceeded for a difference to be significant at some alpha level is given by $t(d)\sqrt{2(MS_w)/n}$. If the group sizes are not equal and the homogeneity of variance assumption is tenable, the use of the harmonic mean for each pair of groups is suggested.

Suppose a study has been run with 17 subjects per group. The error term is 67.24, and the means are as follows:

Control	Treat. 1	Treat. 2	Treat. 3
26.1	29.3	34.2	31.6

Use Dunnett's procedure at the .05 level to determine which of the treatment means is significantly different from the control group mean.

17. The following is an approximate 40% random sample of the CLINICAL data (in Appendix A in back of book), where we present data on only the FREEDIST variable:

GROUP 1	GROUP 2	GROUP 3
9.00	11.00	5.33
9.00	7.67	14.00
7.67	12.00	10.00
8.67	9.00	8.67
14.67	7.33	9.33
6.33	7.33	9.00
7.67	7.33	9.00
7.00	8.33	9.67
7.33	5.33	16.33
8.67	11.00	10.33
8.67	7.67	9.67
7.67	12.00	10.33
8.00	6.00	11.33
12.00	9.67	
7.00	9.33	
5.33	10.00	
8.00	9.67	
9.33		
5.00		
8.33		
9.00		

- (a) Do a one way ANOVA on this data using either SAS or SPSS. What is the null hypothesis? Do you reject at the .10 level?
- (b) Since the group sizes are sharply unequal, test the assumption of equal population variances with the Levene test. Is it significant at the .05 level?
- (c) Apply the Tukey procedure at the .10 level. Which pairs of groups are significantly different?
18. On the following page is a sampling of 60 (all of site 1—three to five year old disadvantaged children from inner city areas in various parts of the country) from the SESAME STREET data base. The grouping variable is VIEWCAT (coded as 1 if the children rarely watched the show to 4 if the children watched the show on average of more than 5 times a week). The dependent variable is POSTLET-PRELET, that is, a measure of how much the children have gained in their knowledge about letters.

VIEWCAT1	VIEWCAT2	VIEWCAT3	VIEWCAT4
7	-1	7	6
3	4	7	6
8	17	28	4
0	6	16	4
4	9	32	24
-1	6	8	27
2	4	-1	21
-1	10	10	-15
-1	9	26	4
	11	-22	24
	1	11	35
	-2	32	5
	6	10	8
	-10	33	7
	21	14	
	5	33	
	7	30	
		31	
		14	
		5	

- (a) Do a one way ANOVA on this data at the .05 level of significance using either SAS or SPSS. What is the null hypothesis? Do you reject it?
- (b) Since the group sizes are sharply unequal, test the assumption of equal population variances with the Levene test. Is it significant at the .05 level? Should we be concerned about the result in (a) being spurious? Explain.
- (c) Apply the Tukey procedure at the .05 level. Which pairs of groups are significantly different?
19. It was mentioned in the chapter that $1 - (1 - \alpha')^k$ is approximately equal to $k\alpha'$ for small α' .
- (a) Let $k = 3$. Expand $1 - (1 - \alpha')^3$, and show that it equals $3\alpha' - 3\alpha'^2 + \alpha'^3$.
- (b) Let $\alpha' = .01$. Calculate $k\alpha'$ and $3\alpha' - 3\alpha'^2 + \alpha'^3$. What have we shown?
20. Run the CLINICAL data set.
- Is there a need to use the Levene test? Explain.
- Are the groups significantly different at the .05 level?

21. Run the ALCOHOL data set.
Is there a need to use the Levene test?
Are the groups significantly different at the .05 level?
Did anything interesting happen in the Tukey post hoc procedure?
22. Why are correlated observations a real problem in social science research?
In answering this question, deal with two issues:
 - (a) How often do correlated observations occur?
 - (b) Examine Table 2.1

APPENDIX

Theorem. The estimated variance for a contrast \hat{L} is given by

$$\hat{\sigma}_{\hat{L}}^2 = MS_w \sum (c_i^2 / n_i),$$

where the n_i , are the groups sizes for the k groups.

Proof. The estimated contrast is

$$\hat{L} = c_1 \bar{x}_1 + c_2 \bar{x}_2 + \cdots + c_k \bar{x}_k$$

Now we take the variance for both sides

$$\text{var}(\hat{L}) = \text{var}(c_1 \bar{x}_1 + c_2 \bar{x}_2 + \cdots + c_k \bar{x}_k)$$

Since the means are sampled from independent groups, the random variables $c_1 \bar{x}_1, \dots, c_k \bar{x}_k$ are uncorrelated. But for uncorrelated variables the variance of a sum is equal to the sum of the variances (since all the covariance terms are 0), Thus,

$$\text{var}(\hat{L}) = \text{var}(c_1 \bar{x}_1) + \text{var}(c_2 \bar{x}_2) + \cdots + \text{var}(c_k \bar{x}_k)$$

Now recall from introductory statistics that if c is a constant, then the variance of cx is $\text{var}(cx) = c^2 \text{var}(x)$.

Using this result the above may be rewritten as

$$\text{var}(\hat{L}) = c_1^2 \text{var}(\bar{x}_1) + c_2^2 \text{var}(\bar{x}_2) + \cdots + c_k^2 \text{var}(\bar{x}_k)$$

Now, it is well known that the variance of a sample mean based on a sample of size n is given by $\text{var}(\bar{x}) = \sigma^2/n$, where σ^2 is the variance of the population. (See Glass & Hopkins, 1984, pp. 188–190, for a proof.) Applying this result to \bar{x}_1, \bar{x}_2 , etc., we obtain:

$$\text{var}(\hat{L}) = c_1^2 \sigma^2 / n_1 + c_2^2 \sigma^2 / n_2 + \cdots + c_k^2 \sigma^2 / n_k$$

In obtaining the above we assumed the population variance was the same in each of the k groups, which is the homogeneity of variance assumption for ANOVA. Now, factoring out the common term σ^2 we have

$$\text{var}(\hat{L}) = \sigma^2(c_1^2 / n_1 + c_2^2 / n_2 + \cdots + c_k^2 / n_k)$$

In practice σ^2 has to be estimated, and recall from the chapter that $MS_w = \hat{\sigma}^2$. Thus, replacing σ^2 by MS_w and writing the sum of the terms in parentheses using the summation operator, we obtain the result stated in the theorem:

$$\text{var}(\hat{L}) = MS_w \sum c_i^2 / n_i$$

Power Analysis

CONTENTS

- 3.1 Introduction
- 3.2 t Test for Independent Samples
- 3.3 A Priori and Post Hoc Estimation of Power
- 3.4 Estimation of Power for One Way Analysis of Variance
- 3.5 A Priori Estimation of Subjects Needed for a Given Power
- 3.6 Ways of Improving Power
- 3.7 Power Estimation on SPSS MANOVA
- 3.8 Summary

3.1 INTRODUCTION

Recall from Chapter 2 that type I error, or the level of significance (α), is the probability of rejecting the null hypothesis when it is true, in effect saying the groups differ when they do not. The α level set by the experimenter is a *subjective* decision, but it is usually set at .05 or .01 by most researchers. The reason for setting α so low, of course, is to minimize the probability of making this error. In statistical inference we can never be sure we have made the correct decision; however, by setting α very low we can control quite effectively the risk of this type of error occurring. As we shall see shortly, though, it is not always wise to set α as low as .05 or .01, especially when the group sizes are less than 20.

There is another type of error that one can make in conducting a statistical test, and this is called a type II error. Type II error (denoted by β) is the probability of accepting H_0 when it is false, i.e., saying the groups don't differ when they do. Note that only one of these errors can occur in a given study. Either we falsely re-

ject H_0 or we falsely accept H_0 . Now, not only can either of these errors occur, but in addition they are inversely related. That is, as we control on type I error, type II error increases. We illustrate the below for a two group problem with 15 subjects per group and a difference between the means of one half a standard deviation. Notice that as we control on α more severely (from .10 to .01), type II error increases fairly sharply (from .37 to .78).

α	β	$1 - \beta$
.10	.37	.63
.05	.52	.48
.01	.78	.22

The quantity in the last column is the *power* of a statistical test, and is the probability of rejecting H_0 when it is false. Thus, power is the probability of making a correct decision. In the above example, if we are willing to take a 10% chance of rejecting H_0 falsely, then we have a 63% chance of finding a difference of a specified magnitude in the population (more specifics on this later in the chapter). On the other hand, if we insist on only a 1% chance of rejecting H_0 falsely, then there are only about 2 chances out of 10 of finding the difference (i.e., power = .22). This example with small sample size suggests that in this case it might be prudent to abandon the traditional α levels of .01 or .05 (and especially .01) to a more liberal α level to improve power sharply. Of course, one does not get something for nothing. We are taking a greater risk of rejecting falsely, but that increased risk is *more than balanced* by the increase in power.

Cast in a broad context, power is dependent on many factors: (1) α level, (2) sample size, (3) effect size, (4) the statistical test used, and (5) the research design. For example, the t test for dependent samples is more powerful than the t test for independent samples, and a repeated measures design (discussed in Chapter 5) is more powerful than a one way ANOVA design. However, the power of a specific statistical test is dependent on these 3 factors:

1. The α level set by the experimenter.
2. Sample size.
3. Effect size—How much of a difference the treatments make, or the extent to which the groups differ in the population on the dependent variable.

We have already indicated that power may be increased substantially by adopting a somewhat more liberal α level, say .10 or .15. There are limits on this, however; no one would take $\alpha = .40$ to gain still greater power, for this is taking far too great a risk of rejecting H_0 falsely.

Power is *heavily* dependent on sample size. Consider a two tailed test at the .05 level for the t test for independent samples. Suppose we have an effect size of .5 standard deviations in the population, i.e., the difference in the means divided by the standard deviation is .5. The table below shows how power changes dramatically as sample size increases from small (10) to large (100):

n (subjects per group)	power
10	.18
20	.33
50	.70
100	.94

With only 10 subjects per group we have only about a 20% chance of detecting this effect size, whereas with 100 subjects per group we are almost certain of detecting the effect (i.e., rejecting the null hypothesis).

As the above example suggests, when sample size is large (say more than 100 subjects per group) power will generally not be an issue. In these cases power will tend to be adequate ($> .70$) to excellent ($> .90$). It is when one is conducting a study where the group sizes are small ($n < 20$), or when one is evaluating a study that had small group size, that it is imperative to be very sensitive to the possibility of a type II error.

The third factor that affects power is effect size. If the effect size is small or medium, then we will see shortly that large group size is needed to detect these effects, i.e., to have adequate power. On the other hand, if the effect size is large (about one standard deviation or greater), then only about 15 subjects per group will be needed for adequate power.

Most statistics books do not do a good job of discussing the *consequences* of making a type I or a type II error. Let me attempt to remedy this deplorable situation. Suppose you are comparing a treatment group vs. a control group on some dependent (outcome) variable. Treatment here is generic and could refer to teaching method, counseling method, diet, drug, etc. Schematically:

TREAT	CONTROL
μ_1	μ_2

The null and alternative hypotheses are as follows:

$$H_0: \mu_1 = \mu_2 \quad H_a: \mu_1 \neq \mu_2$$

If a type I error is made (rejecting H_0 when it is true) we are saying that the treatment is effective when in fact it is not. This is false optimism. A school district, for

example, may invest in a program heavily. If a statistical test is done and a type I error is made, the program is not effective and yet much money has been spent.

Now consider the other side of the coin, that is, a type II error. A type II error is accepting the null hypothesis when it is false. If a type II error is made, we may have “the greatest thing since sliced bread” and not know it. This is false negativism. In a medical sense a type II error could be dangerous or deadly. It would be like telling someone they don’t have a disease when in fact they do. In this case, someone may die before it is realized that a type II error was made.

3.2 *t* TEST FOR INDEPENDENT SAMPLES

Cohen (1977) has defined the population effect size as

$$d = (\mu_1 - \mu_2) / \sigma \quad (1)$$

where σ is the assumed common population standard deviation. This population effect size is estimated by

$$\hat{d} = (\bar{x}_1 - \bar{x}_2) / s$$

where

$$\hat{\sigma}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = s^2$$

is the estimate of the assumed common population variance.

It is necessary to divide by σ in obtaining the effect size measure to adjust for scaling differences on variables which can be quite arbitrary in social science research. Note that Equation 1 expresses the difference between the groups in standard deviation units. For example, if the means for the groups were $\bar{x}_1 = 10$ and $\bar{x}_2 = 4$, with the estimated standard deviation $s = 15$, then $\hat{d} = (10-4)/15 = .4$, or the groups differ by .4 of a standard deviation.

An effect size around .20 is considered small, an effect size around .50, medium, and an effect size $>.80$ is large. A medium effect size is one that would be apparent to a researcher. For example, .5 standard deviations is the difference in mean I.Q. between semiskilled workers and professionals and managers. The difference in mean I.Q.s between PhDs and typical college freshmen is an example of a large effect size, that is, about .8 standard deviations.

The following power values for the *t* test ($\alpha = .05$, two tailed test) from Cohen’s text illustrate precisely how poor power is with small group size and/or small effect size:

1. $n_1 = n_2 = 15$, $d = .50$, power = .26
2. $n_1 = n_2 = 35$, $d = .30$, power = .23

Power can be adequate with small group size, but only if the effect size is *large*. For example, with $n_1 = n_2 = 15$ and $d = 1$, then power = .75 at $\alpha = .05$ for a two tailed test.

Cohen and many others have noted that *small and medium effect sizes are very common in social science research*. Light and Pillimer (1984) in *Summing Up* comment on the fact that most evaluations find small effects in reviews of the literature on programs of various types (social, educational, etc.): “Review after review confirms it and drives it home. Its importance comes from having managers understand that they should not expect large, positive findings to emerge routinely from a single study of a new program. Indeed *any* positive findings are good news” (pp. 153–154). To further document the fact that small and medium effect sizes are common, we present in Table 3.1 the effect sizes for three sets of studies in quite different areas. Note that there are *only 3 large effect sizes out of 40*.

How does one estimate power if the group sizes are unequal? Cohen has suggested using the harmonic mean. Recall from Chapter 2 that the harmonic mean for two groups is given by $2n_1n_2/(n_1 + n_2)$. Thus, if we had group sizes of 10 and 20, we compute the harmonic mean as $2(10)(20)/30 = 13.3$, and use 13 as the n with which to enter Cohen’s power tables. Note that use of the harmonic mean weights the estimate of power down toward the *smaller* group size. The difference between the ordinary mean and harmonic mean is relatively small when group sizes are approximately equal, but when the group sizes are sharply unequal the difference can be considerable, as the following table shows:

Group 1	Group 2	Mean	Harmonic mean
10	15	12.5	12
10	20	15.0	13.3
10	30	20.0	15.0
10	40	25.0	16.0

Researchers not sufficiently sensitive to the power problem may interpret nonsignificant results from studies as demonstrating that “treatments” made no difference. In fact, however, it may be that treatments did make a difference, but that the researchers had poor power for detecting the difference. The poor power may result from small sample size (e.g., < 20 Ss per group) and/or from small effect size. *The danger of low power studies is that they may stifle or cut off further*

TABLE 3.1
Effect Sizes for Three Sets of Studies: Teacher Expectancy, Desegregation,
and Gender Influenceability (Data from Becker, 1987)

<i>Teacher Expectancy</i>				
<i>Study</i>	<i>Sample Size</i>		<i>Effect Size</i>	
	<i>n_e</i> *	<i>n_c</i>		
1	79	339	.03	
2	60	189	.12	
3	72	72	-.14	
4	11	22	1.18	
5	11	22	.26	
6	129	348	-.06	
7	110	636	-.02	
8	26	99	-.32	
9	75	74	.27	
10	32	32	.80	
11	22	22	.54	
12	43	38	.18	
13	24	24	-.02	
14	19	32	.23	
15	80	79	-.18	
16	72	72	-.06	
17	65	255	.30	
18	233	224	.07	
19	65	67	-.07	

<i>Desegregation</i>				<i>Gender Influenceability</i>			
<i>Study</i>	<i>Des**</i>	<i>Seg</i>	<i>Effect</i>	<i>Study</i>	<i>M</i>	<i>F</i>	<i>Effect</i>
1	27	32	.32	1	70	71	-.22
2	39	36	.37	2	60	59	.04
3	28	38	.49	3	118	136	.35
4	29	35	.12	4	77	114	-.30
5	36	35	.22	5	32	32	.63
6	42	35	.29	6	10	10	.81
7	25	48	.59	7	45	45	.39
8	24	48	-.32	8	30	30	.46
9	38	42	-.20	9	40	40	.36
10	131	78	.19	10	61	64	-.06
11	37	101	.24				

**n_e* and *n_c*—numbers of subjects in experimental and control groups

**Des—number of desegregated schools, Seg—number of segregated schools

research in an area where effects do exist, but perhaps are more subtle (as in personality, social, or clinical psychology).

In introductory statistics courses one vs. two tailed tests, or directional vs. non-directional alternative hypotheses, were discussed. It was indicated that one should do a one tail test if there is empirical evidence (previous studies) and/or theory to suggest a difference in a specified direction. The statistical advantage of a one tail test over a two tail test is that it is more powerful. In one of the exercises you are asked to explain why this is so. To further dramatize the considerable difference it can make in power if one adopts a somewhat more liberal α level *and* is able to do a one tail test, consider the table below:

Power of *t* test for independent samples
($n_1 = n_2 = 20$)

a level & nature of test	moderate effect size ($d = .6$)	large effect size ($d = .8$)
$\alpha = .01$, two tail	.22	.44
$\alpha = .05$, one tail	.59	.80
$\alpha = .10$, one tail	.72	.89

At the traditional α level of .01 with a two tail test, power is poor in both cases, while at the .10 level, with the added advantage of a one tail test, power is good in both cases.

3.3 A PRIORI AND POST HOC ESTIMATION OF POWER

If a researcher is going to invest a great amount of time and money in carrying out a study, then he or she would certainly want to have a 70 or 80% chance (i.e., power = .70 or .80) of finding a difference if one is there. Thus, the a priori estimation of power alerts the researcher as to how many subjects per group are needed to have adequate power. This is an important part of experimental planning. More on this shortly.

The post hoc estimation of power is important in terms of how one interprets the results of completed studies. The following example shows how important an awareness of power can be. Cronbach and Snow (1969) had written a report on aptitude-treatment interaction research, not being fully cognizant of the importance of power. By the publication of their text, *Aptitude and Instructional Methods* (1977), on the same topic they acknowledged the importance of power, stating in the preface, “(We) ...became aware of the critical relevance of statistical power,

and consequently changed our interpretations of individual studies and sometimes of whole bodies of literature.” Why would they change their interpretation of a whole body of literature? Because, prior to being sensitive to power, when they found most studies in a given body of literature had nonsignificant results, they concluded no effect existed. However, *after* being sensitized to power they took into account the sample sizes in the studies, and also the magnitude of the effect sizes. If the sample sizes were small in most of the studies with nonsignificant results, then lack of significance is due to poor power. Or, in other words, *several low power studies that report nonsignificant results of the same character are evidence for an effect*. By the same character we mean that the test statistic is “leaning” in the same direction in all cases.

Incidentally, the effect size (\hat{d}) for the t test can be expressed in terms of the t statistic as follows:

$$\hat{d} = t\sqrt{1/n_1 + 1/n_2} \quad (2)$$

where n_1 and n_2 are the respective group sizes. This equation is very helpful in estimating the power of completed studies, since from Equation 2, \hat{d} is quickly computed and then Cohen’s power tables are entered.

3.4 ESTIMATION OF POWER FOR ONE WAY ANALYSIS OF VARIANCE

To define the population effect size for a one way ANOVA we first define a measure of variability of the group means about the grand mean *which is independent of sample size*:

$$\hat{\sigma}_m^2 = \frac{\sum n_i (\bar{x}_i - \bar{x})^2}{N}$$

Where \bar{x} is the grand mean and N is total sample size. Dividing by N makes the measure independent of sample size. Notice that the numerator is just SS_b . To make our effect measure scale free, we again divide by σ as was done for the t test effect measure:

$$f = \sigma_m / \sigma \quad (3)$$

This measure represents the standard deviation of the standardized means, i.e., variability of z score group means about the grand mean. Now, it can be shown (this is one of the exercises) that the estimated effect size can be expressed in terms of the F statistic as follows:

$$\hat{f} = \sqrt{(k-1)F/N} \quad (4)$$

Cohen (1977) characterizes an f around .1 as a small effect size, an f around .25 as medium, and an $f > .4$ as a large effect size. The above equation is quite useful for post hoc estimation of power, since all one needs is the F statistic from the study to obtain the corresponding effect size. With the effect size and the common group size (or the *average* group size if the group sizes are unequal), the power for the study is easily determined using Cohen's power tables. We give two examples to illustrate.

Example 1

A three group study was done by Harrington (1968) on the efficacy of advance organizers in mathematics. He had 10 subjects per group and obtained $F = 4.38$. What was his power at $\alpha = .05$? First, we obtain the estimated effect size using Equation 2:

$$\hat{f} = \sqrt{(k-1)F/N} = \sqrt{2(4.38)/30} = .54 \text{ (large effect size)}$$

In using Cohen's tables recall that n is the common number of subjects per group. Also, for a one way ANOVA, the u in the tables refers to the between groups degrees of freedom, which is $(k-1)$. Thus, here we have $u = 2$. We find that power = .64 at $f = .50$ and power = .81 at $f = .60$ (cf. Table C.2). Therefore, by interpolation, power is estimated as .73, which is adequate. Incidentally, Harrington's F was significant at the .05 level. Note that Harrington had adequate power to reject H_0 in spite of his small sample size because his treatment effect was very large.

Example 2

Consider a four group study with $n_1 = 15$, $n_2 = 13$, $n_3 = 20$, and $n_4 = 17$. The investigator obtained $F = 1.4$. What was her power at $\alpha = .05$? at $\alpha = .10$?

First we obtain the effect size:

$$\hat{f} = \sqrt{2(1.4)/65} = .254 \text{ (medium effect size)}$$

Next, the average group size is 16.25 (we use 16), and $u = k-1 = 3$. Therefore, at $\alpha = .05$, power = .34 (Table C.3), whereas at $\alpha = .10$, power = .48 (Table C.7). In both cases the power is inadequate.

If a post hoc power analysis is done on a study where significance is not found and the effect size is quite small ($< .10$), then one must decide whether such an effect has any practical significance. On the other hand, when significance is not found and a post hoc power analysis reveals a large or medium effect size, then it is essential to replicate the study with more adequate sample size.

3.5 A PRIORI ESTIMATION OF SUBJECTS NEEDED FOR A GIVEN POWER

Here we need an *expected* effect size in order to enter the power tables to determine how many subjects per group are necessary for a specified power at some α level. One could use the average of the estimated effect sizes from studies similar to yours; that is, similar in nature of treatments, duration of treatments, type of subjects, dependent variable used, instrument used to measure the dependent variable, etc. When a study is similar in enough of the above respects to qualify as an estimator of your expected effect size is of course a subjective judgement. But even if an estimate is fairly rough, as long as we can obtain at least two such estimates, the average of these will probably be reasonably accurate. Furthermore, it is surely better to have some estimate and hence be able to determine approximately how many subjects are needed, rather than to have no idea at all.

Example 3

Suppose investigator X has found two studies similar to his.

Study 1: 3 groups, $N = 42$, $F = 2.16$

Study 2: 3 groups, $N = 81$, $F = 1.42$

Now, by using Equation 2 relating F and effect size we find:

$$\hat{f}_1 = \sqrt{2(2.16)/42} = .32 \text{ and } \hat{f}_2 = \sqrt{2(1.42)/81} = .187$$

Thus, the expected effect size for investigator X's study is $(.32 + .187)/2 = .25$. Now, the investigator wishes to know how many subjects per group are necessary to have power = .70 at $\alpha = .05$ and at $\alpha = .10$. Referring to Table C.3 and reading down the column under $f = .25$ until we reach a power value of at least .70, we see that 42 subjects *per group* will be needed at .05 level. Now, using Table C.6 for the .10 level, we see that 32 subjects per group are needed.

In the above example the effect sizes were weighted evenly in determining the expected effect size. However, if one of the studies were much more similar to the study being conducted, then the investigator should weight that effect size more heavily, perhaps giving it double the weight.

3.6 WAYS OF IMPROVING POWER

Given how poor power is generally with less than 20 subjects per group, the investigator should consider the following four ways of improving power:

1. Adopt a more lenient α level, perhaps $\alpha = .10$ or $\alpha = .15$.
2. Use one tailed tests where the literature supports a directional hypothesis.
3. Consider ways of reducing within group variability, so that a more sensitive design results. One way is through sample selection; more homogeneous subjects will tend to vary less on the dependent variable. For example, use just males, rather than males and females, or just use 6 and 7 year old children rather than using 6 through 9 year old children. Another way is through the use of factorial designs, which will be considered in Chapter 4. A third way to reduce within group variability is through the use of analysis of covariance, to be covered in Chapter 7. Covariates that have low correlations with each other are particularly helpful because each is removing a somewhat different part of the within group variance. A fourth way is through the use of repeated measures designs. These designs are very helpful because individual differences due to the average response of subjects are removed from the error term, and such differences are the main reason for within group variability.
4. Make sure there is a strong linkage between the treatments and the dependent variable, and that the treatments extend over a long enough period of time to produce a large or at least a fairly large effect size.

It needs to be mentioned that how far one “pushes” the power issue depends on the *consequences* of making a type I error. One of the reviewers of this text noted that discussing power versus risk reduction (type I error) is most meaningfully considered within a given context and gave the following examples: “If I am testing two teaching methods which cost the same, I go for power. If one method is 10 times more dollars, I go for risk reduction. If I’m comparing drugs A and B and one has some potent side effects, I go for risk reduction.”

In the teaching methods example, if a type I error is made in concluding that the method that is 10 times more expensive is more effective, this will be a very costly mistake for a school district. If a type I error is made in the drug example in concluding drug A (with potent side effects) is better when it is not, this will have serious health consequences for future subjects receiving drug A.

The point the reviewer was making, which is well taken, is that using $\alpha = .10$ or $.15$ to improve power in some cases may not be a wise choice.

3.7 POWER ESTIMATION ON SPSS MANOVA

Starting with Release 2.2, you can obtain power estimates for various statistical tests using the SPSS MANOVA program with the POWER subcommand. To quote from the *SPSS User's Guide* (1988, 3rd edition), "The POWER subcommand requests power valued based on fixed-effects assumptions for all univariate and multivariate F and T tests" (p. 601). Power can be obtained for any α level between 0 and 1, with .05 being the default value. If we wish power at the .05 level for a one way ANOVA, we simply insert the following subcommand: POWER = F(.05)/, or if we wish to know the power at the .15 level: POWER = F(.15)/. We give two examples to illustrate use of the POWER subcommand. The first is the *t* test for independent samples example from Chapter 1 (Section 1.2), and the second is the one way ANOVA example from Chapter 2 (Section 2.9). The command lines for both and selected printout showing the power values are given in Table 3.2.

The effect size measure in each case in Table 3.2 is partial eta-squared (η_p^2), which is given by

$$\eta_p^2 = (df \cdot F) / (df_h \cdot F + df_e) \quad (5)$$

where df_h denotes degrees of freedom for hypothesis and df_e denotes degrees of freedom for error (Cohen, 1973). As the *SPSS User's Guide* (1988, p. 602) notes, "partial eta squared is an overestimate of the actual effect size. However, it is a consistent measure of effect size and is applicable to all *F* and *t* tests."

Cohen (1977, p. 281), in discussing power in one way ANOVA, notes the following relationship between η^2 and the effect size *f*:

$$\eta^2 = f^2 / (1 + f^2) \quad (6)$$

By squaring Equation 4 we find that $\hat{f}^2 = (k-1)F/N$. Plugging this into Equation 6, and with some algebraic simplification, we find that

$$\eta^2 = \frac{(k-1) \cdot F}{(k-1) \cdot F + N} \quad (7)$$

Now, let us compare this with what partial η^2 will be for a one way ANOVA. For one way ANOVA, $df_h = (k-1)$ and $df_e = (N-k)$. Plugging these into Equation 5 we obtain

$$\text{partial } \eta^2 = (k-1) \cdot F / [(k-1) \cdot F + (N-k)] \quad (8)$$

Thus the only difference between η^2 and partial η^2 for one way ANOVA is *N* vs. (*N*—*k*) in the denominator. Since the denominator for Equation 8 will always be smaller than for Equation 7, partial will be an overestimate of effect size; however, for moderately large *N* (say > 50) the difference between the two is small.

TABLE 3.2
Power Analysis Runs on SPSS MANOVA for *t* test for Independent
Samples and for a One Way ANOVA

<i>t</i> test	One Way ANOVA
TITLE 'T TEST FROM CHAP. 1'.	TITLE 'ONE WAY ANOVA FROM 2.8'.
DATA LIST FREE/TREAT PERFORM.	DATA LIST FREE/GPID Y.
BEGIN DATA.	BEGIN DATA.
1 2 1 5 1 5 1 6 1 6 1 7 1 8	1 2 1 3 1 5 1 6
1 9 2 1 2 1 2 2 2 3 2 3 2 4	2 7 2 9 2 1 1
2 5 2 7 2 7 2 8	3 4 3 4 3 5 3 8 3 3
END DATA.	4 8 4 4 4 7 4 7
MANOVA PERFORM BY TREAT (1, 2)/	END DATA.
POWER = F (.05)/	MANOVA Y BY GPID (1,4)/
PRINT = CELLINFO (MEANS)	① POWER = F (.05)/
SIGNIF(EFSIZE)/.	PRINT = CELLINFO (MEANS)
	② SIGNIF (EFSIZE)/.

For Position ONLY

Pick up shaded area from p. 138 of previous edition.

①This is the POWER subcommand to obtain estimated power at the .05 level.

②This subcommand is needed to obtain the effect size measure $\text{partial}\eta^2$.

③Note that power is poor here. This was the example with large effect size but small sample size.

In discussing η^2 for one way ANOVA, Cohen (1977) characterizes $\eta^2 = .01$ as corresponding to a small effect size, $\eta^2 = .06$ to a medium effect size, and $\eta^2 = .14$ to a large effect size.

3.8 SUMMARY

1. Power is the probability of rejecting the null hypothesis when it is false.
2. Power for a specific statistical test is dependent on (a) level of significance, (b) sample size, and (c) effect size. It is important to realize that power is *heavily* dependent on sample size.
3. Small and medium effect sizes are very common in social science research.
4. Cohen has provided the following rough, but useful, guidelines for small, medium, and large effect sizes for the t test and for one way analysis of variance:

$$\begin{aligned} t \text{ test: } \hat{d} &= .20 \text{ (small), } \hat{d} = .50 \text{ (medium), } \hat{d} > .80 \text{ (large)} \\ F \text{ test: } \hat{f} &= .10 \text{ (small), } \hat{f} = .25 \text{ (medium), } \hat{f} > .40 \text{ (large)} \end{aligned}$$

5. Post hoc estimation of power for completed studies is important in properly interpreting results. In particular, lack of significance in small sample studies *may* be simply due to inadequate power.

6. The following relationships exist between the t and F statistics and their corresponding effect size measures:

$$\hat{d} = t\sqrt{1/n_1 + 1/n_2}, \quad \hat{f} = \sqrt{(k-1) \cdot F/N}$$

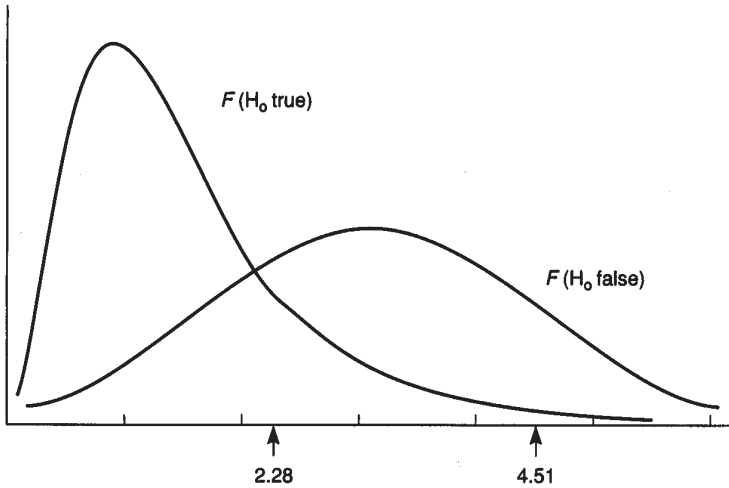
Using these relationships, the corresponding effect size can be easily computed for any study in the literature and then Cohen's power tables entered to determine power.

7. A priori determination of sample size required for a given power at some α level requires an estimate of the anticipated effect size. This estimate can be obtained using estimated effect sizes from previous similar studies and/or from theory.

EXERCISES

1. Graphically, type I error is the area under the F distribution (for H_0 true) in the critical region, while type II error is the area under the F distribution (for H_0 false) *not* in the critical region. Below are given the F distributions for H_0 true and for a case when it is not true for the situation of 4 groups and

30 error degrees of freedom. Also shown are the critical values for $\alpha = .10$ and $.01$.



- Using different degrees of shading, indicate what areas correspond to alpha levels of $.10$ and $.01$.
 - Now, using lining, cross hatching, etc., indicate what areas correspond to type II errors for the above alpha levels.
 - As the alpha level decreases in (a), what happens to the sizes of the areas?
 - As the alpha level decreases, what happens to the sizes of the areas corresponding to type II error? Thus, what have we shown graphically?
- Starting from the following form for the t test for independent samples

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1) \cdot s_1^2 + (n_2 - 1) \cdot s_2^2}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

show that $\hat{d} = t\sqrt{1/n_1 + 1/n_2}$ as given in Equation 2.

- Explain why for the same alpha level a one tail test is more powerful than a two tail test.
- A marketing research study by Sternthal, Dholakia, and Leavitt (1978) was designed to test cognitive response theory predictions about the persuasive

efforts of source credibility and initial opinion on number of counterarguments generated. There were 37 subjects—17 who had a positive prior opinion and 20 who were initially negative. Each of these subjects was assigned to either a moderate or high source credibility condition. As predicted, the moderate credibility source subjects generated more counterarguments; however, the t statistic was not significant ($t = 1.43$, $df = 35$).

- (a) What is the effect size in this study?
- (b) Estimate power at the .05 level.
- (c) What might the investigators consider doing in a future replication study?

5. A researcher in counselor education reviews a small number of studies that have compared (a) counselors in training who participate in a classroom discussion on counseling skills and (b) counselors in training who both participate in a classroom discussion and also observe a videotape of expert counselors outside of the classroom. The dependent measure is empathy. He finds that only one of 10 such studies shows statistical significance at the .05 level. He thus concludes that the effectiveness of the videotape has not been established. Below are the sample sizes and associated t values for the studies:

<i>Study</i>	<i>Classroom Discussion and Videotape</i>	<i>Classroom Discussion</i>	<i>t</i>
1	10	8	1.46
2	12	12	1.76
3	11	13	1.37
4	25	20	1.23
5	6	8	1.64
6	49	36	3.08*
7	20	20	-1.13
8	21	23	1.59
9	8	9	1.92
10	10	10	-.45

The results favor the classroom discussion and videotape group in all cases except studies 7 and 10.

- (a) Calculate the effect size for each of the studies.
- (b) From an examination of the effect sizes, and considering power, might you come to a different conclusion concerning the effectiveness of the combined treatment?

6. Show that $\hat{f} = \sqrt{(k-1) \cdot F / N}$ as given in Equation 4.
7. An ANOVA is run with 5 groups and 25 subjects per group. The F value is 2.03, which is not significant at the .05 level. What is power at the .05 level? at the .10 level?
8. An investigator is in the process of planning a 4 group study in which she will use ANOVA to analyze the results. From previous related literature she estimates that the expected effect size for her study will be .35. How many subjects will she need per group for power = .70 at $\alpha = .05$? At $\alpha = .10$? How many subjects would be needed per group at the same alpha levels if she wanted power to be .80?
9. A survey researcher compares four religious groups on their attitude toward education. The survey is sent out to 1200 subjects, of which 823 eventually responded. Ten items, Likert scaled from 1 to 5, are used to assess attitude. A higher positive score indicates a more positive attitude. There are only 800 usable responses. The Protestants are split into two groups for analysis purposes. The group sizes, along with the means and standard deviations, are given below:

	<i>Protestant 1</i>	<i>Catholic</i>	<i>Jewish</i>	<i>Protestant2</i>
n_i	238	182	130	250
\bar{x}	32.0	33.1	34.0	31.0
s_i	7.09	7.62	7.80	7.49

An analysis of variance on these four groups yields $F = 311.66/55.58 = 5.61$, which is significant at the .001 level.

- (a) Estimate power at the .05 level for the above example.
- (b) Calculate the effect size, and discuss the practical significance issue.
10. (a) Using SPSS MANOVA, obtain estimates of power for the t test example in Table 3.2 at $\alpha = .10$, .15, and .20.
- (b) Does power become adequate for any of the above alpha levels?
- (c) What does the above suggest might be done in small sample studies?
11. Why do I make the statement in the SUMMARY that "In particular, lack of significance in small sample studies may be simply due to inadequate power"?

12. Suppose we had a 2-group study with 10 subjects in Group 1 and 30 subjects in Group 2. Use Cohen's power tables to determine what power would be for a medium effect size at the .05 level.
13. Kazdin (2003, p. 71) notes, "Indeed, a review of medical research for a variety of diseases and conditions revealed that over 25% of the published studies (1975–1990) surveyed revealed no differences between the treatments that were studied. In the majority of these studies power was very weak." Why is this problematic?

Factorial Analysis of Variance

CONTENTS

- 4.1 Introduction
- 4.2 Numerical Calculations for Two Way ANOVA
- 4.3 Balanced and Unbalanced Designs
- 4.4 Higher Order Designs
- 4.5 A Comprehensive Computer Example Using Real Data
- 4.6 Power Analysis
- 4.7 Fixed and Random Factors
- 4.8 Summary
- Appendix Doing a Balanced Two Way ANOVA With a Calculator

4.1 INTRODUCTION

In Chapter 2 we considered the effect of a single independent (grouping) variable on a dependent variable, called one way ANOVA. In this chapter we extend the discussion to examine the effect of two or more independent variables (factors) on some dependent variable, which is called factorial analysis of variance. Often the interest is in whether the additional independent variable moderates or changes the effect of a primary treatment variable. For example, suppose an investigator believes the effect of three treatments on changing attitude toward minorities will vary according to whether the subjects are male or female. That is, the investigator feels treatments will work differently with these subgroups. Since there are two levels for sex and three levels for treatment, we have what is called a 2×3 factorial design. As another example, suppose an educational psychologist has reason to believe from previous research that teaching method 1 will yield highest achievement for urban elementary children while teaching method 3 will work best with rural

children. He is not sure which method works best with suburban children. He can check out these beliefs by setting up a 3×3 factorial design: three levels for location (urban, suburban, and rural) by three teaching methods.

One broad area of research that utilizes factorial designs is called aptitude by treatment interaction (ATI) research. This research is concerned with the effect of any individual difference characteristic of subjects on their response to treatments. The definitive source on this type of research is *Aptitudes and Instructional Methods* by Cronbach and Snow (1977). Aptitude is defined very generally and includes ability, personality, and nontest factors such as social class, ethnic background, sex, etc. Cronbach and Snow discuss numerous ATI studies that have been done in various areas: (1) interactions of abilities with variations in instructional programming, (2) interactions in reading and arithmetic instruction, (3) interactions of abilities with variations in curriculum and instruction, and (4) interactive effects of making instruction less verbal. The reader with interest in any of the above areas will find the Cronbach and Snow book very interesting, critical, and informative. For applied researchers with a clinical orientation, there is an interesting review article by Dance and Neufeld (1988) on ATI research in the clinical setting. Here the focus is on client variables that predict differential treatment responsiveness. They review the literature encompassing cognitive and/or behavioral treatments for anxiety, depression, pain, obesity, and tobacco dependence.

The previous discussion has focused on one experimentally induced factor (treatments) and some individual difference characteristic of subjects that might moderate the effect of the treatments. Factorial ANOVA can be appropriate, however, any time the subjects are cross-classified on two factors and measured on some dependent variable. For example, suppose a survey researcher cross-classified 200 subjects on sex and religion (Catholic, Jewish, and Protestant) and wished to determine whether attitude toward abortion is influenced by sex and religion. She could test this out with a 2×3 (sex \times religion) factorial design.

Advantages of a Two Way Analysis of Variance

A two way design enables us to examine the *joint (interactive)* effect of the independent variables on the dependent variable. We cannot get this information by running two separate one way analyses. An interaction means that the effect one independent variable has on the dependent variable is *not* the same for all levels of the other independent variable. This moderating effect can take two forms:

(a) the degree of superiority changes, but one subgroup always does better than another to illustrate this, consider the following ability by teaching methods design:

	<i>Methods of Teaching</i>		
	T_1	T_2	T_3
High Ability	85	80	76
Low Ability	60	63	68

The numbers in each cell represent the mean achievement for subjects in that cell; that is, 85 is average achievement for high ability subjects under teaching method 1, and so on. Note that the high ability students do better than the low ability for all teaching methods (as we would expect). However, the superiority of the high ability students changes from 25 for T_1 to only 8 for T_3 . Since the order of superiority is maintained, however, this is called an *ordinal* interaction.

(b) The superiority reverses; that is, one treatment is best with one group, but another treatment is better for a different group. A study by Daniels and Stevens (1976) provides an illustration of this more dramatic type of interaction, called a *disordinal* interaction. Using a group of college undergraduates, they considered two types of instruction: (1) a traditional, teacher controlled (lecture type) and (2) a contract for grade plan. The subjects were classified as internally or externally controlled, using Rotter's scale. An internal orientation means that those subjects perceive positive events occur as a consequence of their actions (i.e., they are in control), while external subjects feel that positive and/or negative events occur more because of powerful others, or due to chance or fate. The design and the means for the subjects on an achievement posttest in psychology are given below:

		<i>Instruction</i>	
		Contract for Grade	Teacher Controlled
<i>Locus of Control</i>	Internal	50.52	38.01
	External	36.33	46.22

The moderator variable in this case is locus of control, and it has a substantial effect on the efficacy of an instructional method. When the subject's locus of control is matched to the teaching method (internals with contract for grade and externals with teacher controlled) the subject does quite well in terms of achievement; where there is a mismatch, achievement suffers.

This study also illustrates how a one way design can lead to quite misleading results. Suppose that Daniels and Stevens had just considered the two methods, ignoring locus of control. The means for achievement for the contract for grade plan and for teacher controlled are 43.42 and 42.11, nowhere near significant. The con-

clusion would have been that teaching methods don't make a difference. the factorial study showed, however, that methods definitely do make a difference, a quite positive difference if subject locus of control is matched to teaching methods, and an undesirable effect if there is a mismatch.

A second advantage of factorial designs is that they can lead to more powerful tests by reducing error (within cell) variance. If performance on the dependent variable is related to the individual difference characteristic (the blocking variable), then the reduction can be substantial. Consider the hypothetical sex \times treatment design below:

	T_1	T_2
Males	18, 19, 21 20, 22 (2.5)	17, 16, 16 18, 15 (1.3)
Females	11, 12, 11 13, 14 (1.7)	9, 9, 11 8, 7 (2.2)

Notice that *within* each cell there is very little variability. The within cell variances quantify this, and are given in parentheses. Recall from Chapter 3 that for equal group sizes the error term was just the average of the group variances. In two way ANOVA, for equal cell sizes, the error term is simply the average of the *cell* variances, which here is 1.925. on the other hand, if this had been considered as a two group design, combining males and females together, then the variability is considerably greater, as evidenced by within group (treatment) variances for T_1 and T_2 of 18.766 and 17.6, and a pooled error term for the t test of 18.18.

A third advantage of a factorial design is *economy of subjects*. We only need *half* as many subjects to do a two way ANOVA as would be needed for *two* one way ANOVAs with the same number of levels for each factor. We use a 3×4 (A \times B) factorial design with 10 subjects per cell to illustrate. Here we need a total of $12(10) = 120$ subjects to test whether a and B have a systematic effect on the dependent variable *and* to test for the joint effect of a and B (the interaction effect). To test whether a and B have systematic effects using two one way ANOVAs is going to require 40 Ss per level for a and 30 Ss per level for B, or a total of 240 subjects, as can be seen from the diagram below:

	B				
A	10	10	10	10	40
	10	10	10	10	40
	10	10	10	10	40
	30	30	30	30	

Overview of the Five Major Sections in the Chapter

Since this is a long chapter, we have split it up into five major sections to help the reader organize and see the different main thrusts. The first major section involves a numerical example, where we show how the sums of squares are calculated for the various effects in a two way ANOVA, and how to test each of the effects for significance. The second major section deals with equal and unequal cell size factorial analysis of variance. It is noted that although equal n is desirable, often in practice unequal cell size occurs. Also, we discuss different ways of analyzing unequal n designs and indicate which method should generally be used. The third major section discusses higher order 3 and 4 way ANOVA designs. A three way ANOVA would arise, for example, if we wished to determine whether *both* sex and race moderated the effect of treatments. We would have a treatment by sex by race design. The focus in this chapter for higher order designs is *not* on calculating sums of squares for the various effects, but rather on interpreting effects that are significant. The fourth major section involves a comprehensive computer example that ties together various concepts that were discussed earlier in the chapter. The final major section deals with power analysis for two and three way ANOVA, and the use of SPSS MANOVA for obtaining power estimates is illustrated.

4.2 NUMERICAL CALCULATIONS FOR TWO WAY ANOVA*

Now that the reader has examples of two way ANOVAs in mind and reasons why a two way ANOVA is advantageous, we consider a small data set to illustrate what the hypotheses are that we are testing and how they are tested. In the one way ANOVA there were just two sources of variation (between and within). In a two way ANOVA ($A \times B$ design) there are four sources of variation:

1. Variation due to factor A .
2. Variation due to factor B .
3. Variation due to the interactive effect of A and B .
4. Within cell (error) variation.

Consider the following 2×3 design with 3 observations per cell.

*We again have the same assumptions as for a one way ANOVA, except now they apply to cells, that is, normality on the dependent variable in each cell and equal cell population variances.

		Treatments (B)			
		1	2	3	Row Means
Sex (A)	1 (males)	12, 16, 17 $\bar{x}_{11} = 15$	13, 9, 8 $\bar{x}_{12} = 10$	14, 15, 13 $\bar{x}_{13} = 14$	$\bar{x}_{1.} = 13$
	2 (females)	6, 10, 8 $\bar{x}_{21} = 8$	11, 8, 8 $\bar{x}_{22} = 9$	12, 10, 8 $\bar{x}_{23} = 10$	$\bar{x}_{2.} = 9$
	Column Means	$\bar{x}_{.1} = 11.5$	$\bar{x}_{.2} = 9.5$	$\bar{x}_{.3} = 12$	$\bar{x} = 11$ grand menu

The first number in the subscript for each cell mean refers to the level for Sex, while the second number refers to the level for Treatment. Thus, \bar{x}_{12} refers to the mean for males in treatment 2, while \bar{x}_{23} refers to the mean for females in treatment 3. The dot notation in the second part of the subscript means we are summing across the columns or levels of factor *B* in obtaining the row means. The dot notation in the first part of the subscript indicates we are summing across the rows or levels of factor *A* to obtain the column means.

As for the one way ANOVA, we will use *definitional* formulas to compute the sums of squares for factor *A*, factor *B*, interaction and error, that is, SS_A , SS_B , SS_{AB} and SS_w . Before we give these formulas, let us have clearly in mind what hypotheses are being tested. The first two involve what are called *main effects* for factors *A* and *B*. The null hypothesis being tested for the *A* main effect is:

$$H_0 : \mu_{1.} = \mu_{2.} \cdots = \mu_{I.} \text{ (population row means are equal)}$$

In the above table $\bar{x}_{1.} = 13 = \hat{\mu}_{1.}$ and $\bar{x}_{2.} = 9 = \hat{\mu}_{2.}$ are estimates of the population means for males and females, and the inferential question is, “Are the differences in the sample row means large enough, given sampling error, to suggest that the underlying population row means are different?” Also, *I* in the above null hypothesis refers to the general number of levels for factor *A*.

The null hypothesis for the *B* main effect is:

$$H_0 : \mu_{.1} = \mu_{.2} \cdots = \mu_{.J} \text{ (population column means are equal)}$$

In the above data table $\bar{x}_{.1} = 11.5 = \hat{\mu}_{.1}$, $\bar{x}_{.2} = 9.5 = \hat{\mu}_{.2}$, and $\bar{x}_{.3} = 12 = \hat{\mu}_{.3}$ are estimates of the underlying population column means. Also, *J* refers to the general number of levels for factor *B*.

Thus, in general we are talking about an $I \times J$ design. We further restrict matters, at this point in the chapter, to what are called *balanced* designs, which are designs with an equal number of observations (n) per cell.

Sums of Squares for Factor A and Factor B

The definitional formula for sum of squares for factor A (SS_A) is given by:

$$\begin{aligned} SS_A &= nJ \sum (\bar{x}_{i.} - \bar{x})^2 \\ &= nJ [(\bar{x}_{1.} - \bar{x})^2 + (\bar{x}_{2.} - \bar{x})^2 + \cdots + (\bar{x}_{I.} - \bar{x})^2] \end{aligned} \quad (1)$$

Note that nJ is the number of observations on which each row mean is based. Thus, this sum of squares merely reflects variability of the row means about the grand mean. It is analogous to the sum of squares between in the one way ANOVA. For our example we have

$$SS_A = 3(3)[(13 - 11)^2 + (9 - 11)^2] = 72$$

We want the mean sum of squares for factor A (MS_A), which is given by

$$MS_A = SS_A / (I - 1) = 72 / 1 = 72$$

The definitional formula for the sum of squares for factor B (SS_B) is given by

$$\begin{aligned} SS_B &= nI \sum (\bar{x}_{.j} - \bar{x})^2 \\ &= nI [(\bar{x}_{.1} - \bar{x})^2 + (\bar{x}_{.2} - \bar{x})^2 + \cdots + (\bar{x}_{.J} - \bar{x})^2] \end{aligned} \quad (2)$$

This sum reflects variability of the column means about the grand mean. For our example it is

$$\begin{aligned} SS_B &= 3(2)[(11.5 - 11)^2 + (9.5 - 11)^2 + (12 - 11)^2] \\ &= 21 \end{aligned}$$

We want mean sum of squares for factor B (MS_B), which is

$$MS_B = SS_B / (J - 1) = 21 / 2 = 10.5$$

Error Term

To test each of these main effects for significance we need an error term. That error term is a pooled within cell measure of variability. Verbally, for each cell we devi-

ate the scores in the cell about the mean for the cell, square the deviations, and add them up across all the cells. Recall that exactly the same process was followed in obtaining the error term for the one way ANOVA, except that the scores were deviated about the *group* means. In symbols we can write the factorial error term as:

$$\sum_{\text{cells}} (x - \bar{x}_{ij})^2 \quad (3)$$

$$SS_w = \underbrace{\sum (x - \bar{x}_{11})^2}_{\text{variability within cell 11}} + \underbrace{\sum (x - \bar{x}_{12})^2}_{\text{variability within cell 12}} + \cdots + \underbrace{\sum (x - \bar{x}_{IJ})^2}_{\text{variability within cell IJ}}$$

Now we compute this for the example:

$$\begin{aligned} SS_w &= (12-15)^2 + (16-15)^2 + (17-15)^2 && \text{(cell 11)} \\ &\quad + (13-10)^2 + (9-10)^2 + (8-10)^2 && \text{(cell 12)} \\ &\quad + (14-14)^2 + (15-14)^2 + (13-14)^2 && \text{(cell 13)} \\ &\quad + \cdots + (12-10)^2 + (10-10)^2 + (8-10)^2 && \text{(cell 23)} \\ SS_w &= 52 \end{aligned}$$

We want MS_w , which, as mentioned earlier, represents the average of the cell variances. This is given by

$$MS_w = SS_w / (N - IJ)$$

A degree of freedom is lost in estimating each cell mean, hence the degrees of freedom for error is $N - IJ$. For the example, we have

$$MS_w = 52 / (18 - 6) = 4.33$$

F Tests for the Main Effects

The *F* tests for the main effects are analogous to doing two one way ANOVAs on the data, although the error term here is different. One can think of “slicing the data cake” first horizontally, and then vertically. The *F* ratio for the a main effect is given by

$$F_A = MS_A / MS_w$$

which for our data becomes

$$F_A = 72 / 4.33 = 16.63$$

The critical value at the .05 level is given by

$$F_{.05;I-1,N-IJ} = F_{.05;1,12} = 4.75$$

Because the value of the test statistic is greater than the critical value (i.e., $16.63 > 4.75$), we reject and conclude that males did significantly better than females.

The F ratio for the B main effect is given by:

$$F_B = MS_B / MS_w ,$$

which for this data is

$$F_B = 10.5 / 4.33 = 2.42$$

The critical value at the .05 level is given by

$$F_{.05;J-1,N-IJ} = F_{.05;2,12} = 3.89$$

Since $2.42 < 3.89$, we fail to reject and conclude that treatments did not have a differential effect on performance. Or, to put it another way, the sample column means are estimating equal population column means.*

The Interaction Effect

We will define the interaction sum of squares (SS_{AB}) in terms of the *cell* interaction effects. A cell interaction effect, which we denote by ϕ_{ij} , is that part of the cell mean that can *not* be accounted for by overall effect (grand mean) and by main effects for a and B . The main effects for a and B respectively are: $\alpha_i = \mu_i - \mu$ and $\beta_j = \mu_{.j} - \mu$, where μ is the grand (overall) population mean. Thus, the cell interaction effect is:

$$\begin{aligned}\phi_{ij} &= (\mu_{ij} - \mu) - (\mu_i - \mu) + (\mu_{.j} - \mu) \\ &= \mu_{ij} - \mu_i - \mu_{.j} + \mu\end{aligned}$$

Now, the sum of squares for interaction is

$$SS_{AB} = n \sum \hat{\phi}_{ij}^2 ,$$

*The reader needs to understand that although we are doing the tests on main effects first because of their simplicity, if a significant interaction effect is found then interpretation of the results needs to be focused on the interaction. An interaction effect means that the explanation of the data requires a more complex model.

where

$$\hat{\phi}_{ij} = \bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x} \quad (4)$$

is the estimated cell interaction effect.

Let us bring back the data again, but this time just with the row and column means and grand mean:

(B) Treatments

		1	2	3	Row Means
Sex (A)	1	$\bar{x}_{11} = 15$ $\hat{\phi}_{11} = 1.5$	$\bar{x}_{12} = 10$ $\hat{\phi}_{12} = -1.5$	$\bar{x}_{13} = 14$ $\hat{\phi}_{13} = 0$	13
	2	$\bar{x}_{21} = 8$ $\hat{\phi}_{21} = -1.5$	$\bar{x}_{22} = 9$ $\hat{\phi}_{22} = 1.5$	$\bar{x}_{23} = 10$ $\hat{\phi}_{23} = 0$	9
	Column Means	11.5	9.5	12	11
					↓ grand mean

The estimated cell interaction effects above were obtained using Equation 4. To illustrate, we calculate the first two:

$$\hat{\phi}_{11} = 15 - 13 - 11.5 + 11 = 1.5$$

and

$$\hat{\phi}_{12} = 10 - 13 - 9.5 + 11 = -1.5$$

It can be shown that for a fixed effects design the sum of the interaction effects for *every* row and column add to 0. Thus, for the above design, once the above two effects are calculated the others are determined. For example, since the sum of the interaction effects for row 1 must be 0, we have

$$1.5 + (-1.5) + x = 0 \text{ or } x = 0$$

Similarly, since the sum of the interaction effects for column 1 must be 0, it follows that $\hat{\phi}_{21} = -1.5$.

Plugging the interaction effects into the equation above:

$$\begin{aligned} SS_{AB} &= 3[(1.5)^2 + (-1.5)^2 + (-1.5)^2 + (1.5)^2] \\ SS_{AB} &= 3(9) = 27 \end{aligned}$$

Now, the mean sum of squares (MS_{AB}) is given by

$$MS_{AB} = SS_{AB} / (I - 1)(J - 1)$$

where $(I - 1)(J - 1)$ is the degrees of freedom for interaction. For this problem we have

$$MS_{AB} = 27 / (2 - 1)(3 - 1) = 13.5$$

The F ratio for interaction is

$$F_{AB} = MS_{AB} / MS_w = 13.5 / 4.33 = 3.12$$

The critical value at the .05 level is given by

$$F_{.05; (I-1)(J-1), N-12} = F_{.05; 2, 12} = 3.89$$

Since $3.12 < 3.89$, the interaction effect is not significant (the null hypothesis is H_0 : All $\phi_{ij} = 0$).

Another way of characterizing an interaction effect is as “a difference in the differences.” the differences across sex for the different treatments are respectively 7, 1, and 4. Although these differences may appear to be large, the test statistic indicates that they are likely to occur from populations with equal differences. This is due to considerably sampling error present here because of the very small cell size of 3.*

Graphically, if there is no interaction, then the *population profile of means will be parallel*. When plotting real data, however, the profiles of sample means will essentially always be non-parallel. For the previous example the interaction effect was not significant at the .05 level, and yet when the profiles of means are plotted (see Figure 4.1) for A_1 and A_2 they are strikingly non-parallel. For this example there was a power problem due to small cell size (3), because the estimated interaction effect size is very large: $f = \sqrt{(2 - 1)(3 - 1)2.12} / 18 = .59$.

Still another example to illustrate that the sample mean profiles will be non-parallel even when the interaction F is less than 1 is provided in later in this section (where we consider three way ANOVA). The social class by grade level interaction F was .868. When the profiles of means were plotted (Figure 4.1) for social class they even cross, although just slightly. The F test is telling us, however, that these sample mean profiles are estimating parallel *population* profiles.

Finally, note that the degrees of freedom for interaction, $(I - 1)(J - 1)$, is directly related to the fact that the sum of the interaction effects for every row and column must add to 0. Recall from our example, once we had computed the first two inter-

*In the Appendix we illustrate how to do an equal n ANOVA using a calculator.

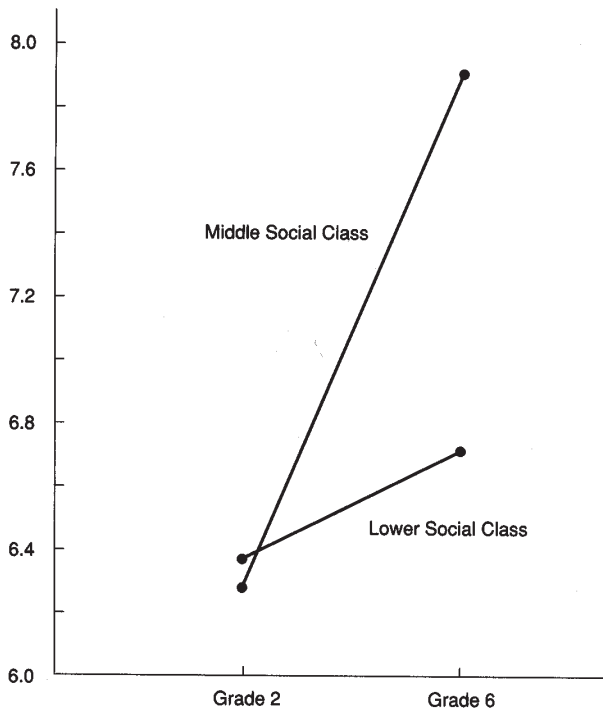
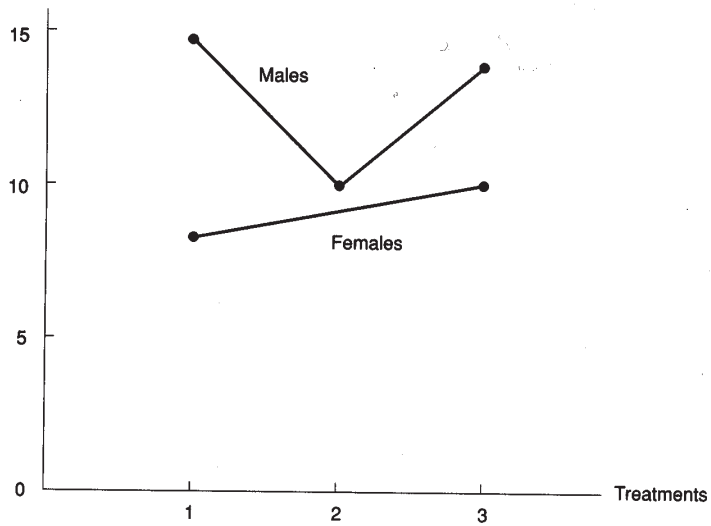


FIGURE 4.1 Interaction Profiles for Two Data Sets

action effects, the others were determined. That is, they were *not* free to vary. This was because for the 2×3 design there are only 2 degrees of freedom.

Linear Model for the Data

Recall that the linear model for a subject's score (y_{ij}) in one way ANOVA was

$$y_{ij} = \mu + \alpha_j + e_{ij} ,$$

where μ was the grand mean (general effect), $\alpha_j = \mu - \mu_j$ was an effect unique to the j th treatment, and e_{ij} was random error. The subjects score was decomposed into three components. In two way ANOVA we still have a linear model, although now there will be a component for factor A , a component for factor B , and a component to represent the joint effect of A and B . The model looks like this:

$$y_{ijk} = \underbrace{\mu}_{\text{general effect}} + \underbrace{\alpha_i + \beta_j}_{\text{main effects}} + \underbrace{\phi_{ij}}_{\text{interaction effect}} + e_{ijk}$$

where $\alpha_i = \mu_i - \mu$ (deviation of i th row mean from grand mean), $\beta_j = \mu_j - \mu$ (deviation of the j th column mean from the grand mean), and ϕ_{ij} is the interaction effect, which was defined in the previous section. The triple subscript for the subject's score is read, "the score for subject k in cell ij ."

Now we show for a few selected subjects from the example in the section on Interaction Effect how their scores can be decomposed into the four parts given by the linear model for a two way ANOVA design. We start with the first subject in cell 11. That score can be expressed as:

$$y_{ijk} = \underbrace{\mu}_{\text{general effect}} + \underbrace{\alpha_i + \beta_j}_{\text{main effects}} + \underbrace{\phi_{ij}}_{\text{interaction effect}} + e_{ijk} \quad (5)$$

Thus, the error component for this subject is -3 .

Or, consider the second subject in cell 21. That score of 8 can be decomposed as follows:

$$8 = \underbrace{11}_{\text{general effect}} + \underbrace{(9-11)}_{\text{A main effect}} + \underbrace{(9.5-11)}_{\text{B main effect}} + \underbrace{(1.5)}_{\text{interaction effect}} + e_{111}$$

Therefore, the error component for this subject is -1 .

4.3 BALANCED AND UNBALANCED DESIGNS

Balanced Designs

In one way ANOVA the total sum of squares was partitioned into two independent sources of variation (sum of squares between and within). In a two way ANOVA, for *equal cell size*, the sums of squares for the main effects, interaction, and error are also independent. That is, SS_A , SS_B , SS_{AB} , and SS_W are independent. Of course, the corresponding mean squares will also be independent. But the F ratios for A , B , and AB interaction are *not* independent. Why? Because they all share the *same* error term, that is, MS_W . However, research has shown that if the total N is even moderately large, then the amount of dependence will be small and can be ignored for practical purposes. Thus, we will regard the F tests as independent. This is important in terms of clarity of interpretation, since significance on one effect is not dependent on (or confounded with) significance on the other effects. We see later on in this chapter that for disproportional cell sizes the effects are correlated or confounded.

Factorial ANOVA on SAS and SPSS

Now we show how easy it is to run the data from the numerical example on SAS and SPSS. The control lines, along with annotation, are given in Table 4.1. Annotated printout from SAS is presented in Table 4.2 and printout for SPSS is given below:

Placekeeper for shaded table from p. 158 of previous edition.

Unbalanced Designs

For the equal cell size designs the sums of squares for the different effects are uncorrelated, and for moderately large N the F ratios for the effects are essentially uncorrelated. This is important in interpreting results since significance on one ef-

TABLE 4.1
SAS and SPSS Control Lines for 2 × 3 Factorial ANOVA

SAS	SPSS GLM
TITLE 'TWO WAY ANOVA';	TITLE 'TWO WAY ANOVA—P. 159'.
DATA TWOWAY;	DATA LIST FREE/FACA FACB DEP.
① INPUT FACA FACB DEP @@;	BEGIN DATA.
LINES;	1 1 12 1 1 16 1 1 17
② 1 1 12 1 1 16 1 1 17	1 2 13 1 2 9 1 2 8
1 2 13 1 2 9 1 2 8	1 3 14 1 3 15 1 3 13
1 3 14 1 3 15 1 3 13	2 1 6 2 1 10 2 1 8
2 1 6 2 1 10 2 1 8	2 2 11 2 2 8 2 2 8
2 2 11 2 2 8 2 2 8	2 3 12 2 3 10 2 3 8
2 3 12 2 3 10 2 3 8	END DATA.
PROC PRINT;	LIST.
PROC GLM;	⑥ UNIANOVA DEP BY FACA
③ CLASS FACA FACB;	FACB/DESIGN/.
④ MEANS FACA FACB FACA*FACB;	
⑤ MODEL DEP = FACA FACB	
FACA*FACB;	

① Recall that the @@ is necessary in order to put the data for more than one subject on the same data line.

② the first two numbers of each triple here are for the cell ID for the subject. Thus, the first triple here indicates the first subject in cell 11 has a score of 12 on the dependent variable, the second triple that the score for the next subject in cell 11 is 16, and the first triple for the *fourth* data line that the score for the first subject in cell 21 is 6.

③ This CLASS statement indicates which of the variables in the INPUT statement are the grouping variables.

④ This MEANS statement is required to obtain the level (row and column) means and the cell means.

⑤ in the MODEL statement we put the dependent variable(s) on the left side and the effects in the design on the right side. Since we wish to test the full factorial model we put in the main effects and the interaction. Observe that the interaction effect is indicated by placing an * between the factors.

⑥ We are using the GLM (general linear model) program, which does both regression analysis and ANOVA for *one* dependent variable. That is why the code name is UNIANOVA, for ANOVA on one dependent variable.

fect implies nothing about significance on another. This makes for a clean and clear interpretation of results. However, often in real world data situations we will not have equal cell size, for at least two reasons:

1. Even if we started with equal cell size in an experimental study, because of experimental mortality (subjects dropping out of the study for various reasons—

parents moving, boredom, annoyance with the treatment, treatment schedule becomes inconvenient, etc.) we wind up with unequal cell sizes.

2. We are studying intact groups, which when cross-classified produce quite different subgroup (cell) sizes. of course, we could in some instances simply randomly discard subjects from cells to achieve equal n , but in other cases this may cause a loss of too many subjects.

Thus it becomes imperative to be able to analyze and properly interpret unequal cell size factorial designs. The problem with disproportional cell size designs is that the effects become correlated (confounded), and unless these correlations are taken into account we may misinterpret the results. There is a considerable amount of literature on the topic, particularly from the late 1960s through the 1970s. Overall and Spiegel (1969), in a classic paper on analyzing factorial designs, discuss three basic methods of analysis:

Method 1: Adjust each effect for all other effects in the design to obtain its unique contribution (regression approach).

Method 2: Estimate the main effects ignoring the interaction, but estimate the interaction effect adjusting for the main effects (experimental method).

Method 3: Based on theory and/or previous research, establish an ordering for the effects, and then adjust each effect only for those effects preceding it in the ordering (hierarchical approach).

For equal cell size designs all three of the above methods yield the same results, that is, the same F tests. Therefore, it will not make any difference, in terms of the conclusions a researcher draws, which of these methods is used on one of the packages. For unequal cell sizes, however, these methods can yield quite different results.

Two Examples for Unbalanced Designs

We give two examples for unequal n factorial designs. The first example uses artificial data, but shows that the method of analysis can affect the conclusions drawn. The second example uses real data, and also illustrates bringing data in from disk. For our first example, consider the following 2×3 design:

		B		
A		3, 5, 6	2, 4, 8	11, 7, 8, 6, 9
		9, 14, 5, 11	6, 7, 7, 8, 10, 5, 6	9, 8, 10

TABLE 4.2
 Factorial ANOVA run for Numerical Example on SAS General Linear Models (GLM) Procedure

Placeholder for table on p. 160 of previous edition

TABLE 4.2 (Continued)

Placeholder for table p.161 of previous edition

① The model sum of squares is in the total of the sums of squares for the effects in the model. Here $120 = 72$ (ss for FACA) + 21 (ss for FACB) + 27 (ss for FACA*FACB).

② Notice from the tail probabilities to the right of the F ratios, that only FACA is significant at the .05 level, since only that probability is less than .05.

③ The mean square error (4.333), which we denoted by MS_w in the chapter, is the denominator for each of the F ratios.

④ Row means for factor A.

⑤ Column means for factor B.

The control lines for running the analysis on SPSS and on SAS are the same as for the equal cell n case. With both programs the regression approach (Method 1) is the default option; that is, it is used automatically unless something else is specified. In both programs the unique sum of squares (regression approach) is called type III sum of squares. Type I sum of squares in both programs refers to the sequential sum of squares. In the sequential approach (also called hierarchical) a given effect is adjusted for all effects to its left (or preceding it) in the ordering. Suppose the effects went in the following order: FACA, FACB, FACA * FACB.

Then, in the sequential approach, the a main effect is not adjusted for anything. The B main effect is adjusted for the a main effect, and the interaction is adjusted for both main effects. In this approach the sums of squares for the terms in the model do add up to the total sum of squares.

We ran the above data on SAS GLM and on the SPSS GLM program (*SPSS Base*, p. 239). Both sums of squares come out in one run on SAS; recall that type III sum of squares is the unique sum of squares, while type I is the sequential sum of squares. Two runs are required for SPSS, although we just present the type III (unique) sum of squares for the SPSS run in Table 4.3.

If we use the unique sum of squares approach we would conclude that only the factor a main effect is significant at the .05 level of significance, because only that p value is less than .05. on the other hand, with the sequential sum of squares the conclusion would be that both main effects are significant at the .05 level (p values of .048 and .043, respectively). Thus, the method used with disproportional designs can make a difference in terms of the conclusions drawn from an experiment. Importantly, however, the interaction F is the *same* for both approaches, because all other effects (main effects) are partialled from it with the unique sum of squares approach, and also both main effects are partialled in the sequential approach since the interaction effect is last in the ordering.

Our research example involves data from a study by Philips and Jahanshahi of the London University Institute of Psychiatry (Hand & Taylor, 1987). The study examined the effectiveness of different kinds of psychological treatment on the sensitivity of headache sufferers to noise. Each subject was first pretested on sensitivity, then given relaxation training (to be defined shortly), then given one of 4 treatments, and finally posttested on sensitivity. The sensitivity scores were obtained by listening to a tone that gradually increased in volume and having the subjects rate the levels at which the tone became (1) uncomfortable and (2) definitely unpleasant. These ratings are the dependent variables for the study. We denote the pretest and posttest ratings by PREU, PREUP, POSTU, and POSTUP.

1. The subjects were asked to listen to the tone at their definitely unpleasant level for up to two minutes, with the option of terminating the exposure if they wished.

2. The subjects were then given instruction on breathing techniques and the use of visual imagery to act as a controlled distraction.

The design was a 2×4 factorial because there were two types of headache sufferers involved (migraine and tension) and four treatments, which were as follows:

T_1 : Subjects in this group listened to the tone again at their initial definitely unpleasant (POSTUP) level for the length of time that they were able to stand it in the relaxation training phase.

T_2 : This treatment was the same as T_1 , but with one extra minute's exposure to the tone.

T_3 : The subjects in this treatment group had the same exposure to the tone as those in treatment group 2, but they were *instructed* to use the relaxation techniques of breathing and visual imagery.

T_4 : This was a control group, in that the subjects had no exposure to the tone between the relaxation training and the posttest measurement.

From within migraine and tension, the subjects were randomly assigned to the treatment groups. However, missing data reduced an initial balanced design to the following unequal n situation:

	T_1	T_2	T_3	T_4
Migraine	11	11	12	11
Tension	14	11	16	12

Here we consider analysis on only the definitely unpleasant posttest rating (DEFUNPL) of the subjects. The raw data for this study is given in Appendix a in the back of the book. I had the data on a 3.5-inch disk, along with several other data sets, and ran the analysis using the GLM program of SPSS for Windows 12.0. In Table 4.4 I present selected printout from that run.

For those who are interested, Stevens (1996, pp. 294–301) shows through dummy coding of the effects that the effects are indeed uncorrelated for balanced designs and correlated for disproportional designs.

Which Method Should Be Used?

After much debate in the statistical literature in the 1970s, there seem now to be a consensus that Method 1 (obtaining the unique sum of squares for each effect) *generally should be used*. For example, this is what Carlson and Timm (1974) recommend, and what Myers (1979) recommends for experimental studies (random assignment involved), or, as he puts it, “whenever variations in cell frequencies can reasonably be assumed due to chance” (p. 403).

TABLE 4.3
Factorial Unequal n Printouts From SAS GLM and SPSS GLM Program

Placeholder for art p. 164 of previous edition.

① Note, as mentioned in the chapter, that the sequential sums of squares from SPSSX are the same as the type I sums of squares from SAS.
 ② Unique sums of squares from SPSSX are the same as type III sums of squares from SAS.

TABLE 4.4
Selected Printout From SPSS GLM for Headache Data

Placeholder for art p. 165 of previous edition
--

When an a priori ordering of the effects can be established, then Method 3 (hierarchical or sequential sum of squares) makes sense. Pedhazur (1982) gives the following example. There is a 2×2 design in which one of the classification variables is race (black or white) and the other classification variable is education (high school or college). The dependent variable is income. In this case one can argue that race affects one's level of education, but obviously not vice versa. Thus, it makes sense to enter race first to determine its effect on income, then to enter education to determine how much it adds in predicting income. Finally, the race \times education interaction is entered.

4.4 HIGHER ORDER DESIGNS

Three Way Analysis of Variance

Here we are examining the effect of three independent variables or factors on some dependent variable. We present three examples to illustrate:

1. An instructional technologist wishes to determine whether teaching method (2), teacher (2), and sex of the child each have an effect on achievement in reading. She has a $2 \times 2 \times 2$ factorial design. This design enables her to determine whether all 3 factors jointly affect achievement in some unique way. For example, perhaps method 1 is particularly effective with teacher 1 working with girls, while method 2 is not effective with teacher 2 working with boys.
2. Consider again the aptitude treatment interaction study by Daniels and Stevens (1976) mentioned earlier. That study examined the effect of locus of con-

trol and teaching method on achievement in an introductory psychology course, and found a disordinal interaction. Internals did better with the contract for grade plan while externals did better with the teacher controlled method of instruction. As a heuristic followup to their study, Daniels and Stevens broke the subjects down into males and females and ran a sex by locus of control by method ANOVA to determine whether the nature of the interaction might be different for males and females (it was not).

3. A study by Marwit and Neumann (1974) provides another illustration of a three way ANOVA. Two black and two white examiners administered standard and nonstandard English forms of the California Reading Test to 60 black and 53 white second graders from a St. Louis public school. Here the race of the subject is one factor, the race of the examiner the second factor, and the format the third factor, while achievement is the dependent variable. The design schematically is this:

		Format	
		Standard English	Nonstandard English
Subject	Examiner		
	Black		
Black	White		
	Black		
White	White		

Recall that in a one way ANOVA there were two sources of variation (between and within), in a two way ANOVA there were four sources of variation (factor *A*, factor *B*, interaction of *A* and *B*, and within cell or error variation), and three hypotheses that were tested: *A* and *B* main effects and interaction effect. How many sources of variation are there in a three way design? the number of sources of variation in general for a *k* way factorial design is 2^k . Thus, for a 3 way design there are $2^3 = 8$ sources of variation, while for a 4 way ANOVA there are $2^4 = 16$ sources of variation. Consider the methods \times teacher \times sex design again. The sources of variation are

METHOD (A)	}	MAIN EFFECTS
TEACHER (B)		
SEX (C)		
METHOD \times TEACHER	}	FIRST ORDER INTERACTIONS
METHOD \times SEX		
TEACHER \times SEX		
METHOD \times TEACHER \times SEX		
WITHIN CELLS (ERROR)		

Each of the 7 effects in the design is tested against the *same* error term, that is, within cells variability (MS_w). Thus, the F ratios would look like this: $F_A = MS_A/MS_w$, $F_B = MS_B/MS_w$, ..., $F_{BC} = MS_{BC}/MS_w$, $F_{ABC} = MS_{ABC}/MS_w$. The process of computing SS_w is exactly the same as for the two way ANOVA, that is, deviate the scores about the means in each cell, square the deviations and then add the squared deviations. The degrees of freedom for error for the two way ANOVA was $N - IJ$ (total number of subjects – number of cells). If we denote the number of levels for the factors in a 3 way ANOVA by I , J , and K , then the degrees of freedom for error in a 3 way is $N - IJK$ (again, the total number of subjects – number of cells).

The main effects involve comparing level means, analogous to comparing row and column means for the two way ANOVA. The first order (or two way) interactions are assessed by examining the pattern of means for the two factors combined over the third factor. For example, the method \times teacher interaction is assessed by examining the means for those two factors with boys and girls combined together. Finally, the three way interaction is going to tell us whether the patterns of means for any two factors differs across the levels of the third factor.

Interpretation of Effects for An Example from Literature

To make this more concrete we consider data from a study by Cradler and Goodwin (1971). They were interested in comparing the way in which three types of reinforcement affected children’s ability to use the word *they* when making up sentences. The children were randomly assigned to three groups: (1) material reinforcement condition—subjects received an M&M candy immediately after using the word *they* at the beginning of a sentence; (2) praise reinforcement—the children were reinforced by the experimenter’s saying “good”; and (3) symbolic reinforcement—the children were simply given a plus mark. The investigators were also interested in whether the reinforcements worked differently for middle and lower class children (second factor in the design), and for different aged children (2nd and 6th graders—the third factor in the design). Below are the means (M) and standard deviations (SD):

Social Class	Grade Level					
	Mat.	Grade 2 Praise	Symb.	Mat.	Grade 6 Praise	Symb.
Middle						
M	5.66	6.64	6.58	5.75	8.25	9.66
SD	2.32	3.28	2.98	1.63	4.12	3.37
Lower						
M	8.41	5.41	5.25	6.75	7.00	6.33
SD	4.23	3.63	2.74	4.20	4.16	3.22

There were 12 subjects in each of the cells. The following ANOVA table was obtained:

Source	<i>df</i>	<i>MS</i>	<i>F</i>
Social Class (A)	1	12.250	.957
Grade Level (B)	1	25.000	1.954
Type of Reinforce (C)	2	1.465	.114
<i>A</i> × <i>B</i>	1	11.111	.868
<i>A</i> × <i>C</i>	2	41.646	3.255*
<i>B</i> × <i>C</i>	2	47.396	3.704*
<i>A</i> × <i>B</i> × <i>C</i>	2	5.298	.414
Error	132	12.792	

Note that none of the main effects is significant, although grade level is closest. Recall that in three way ANOVA main effects test whether the underlying *population* level means are different for the factor under consideration. The level mean for grade 2 is obtained by adding up all the means for grade 2 and dividing by 6:

$$(5.66 + 6.64 + 6.58 + 8.41 + 5.41 + 5.25)/6 = 6.325$$

and similarly for grade 6.

The reinforcement level means are obtained by adding the 4 means for each reinforcement condition and then dividing by 4. Thus, for material, we have

$$(5.66 + 8.41 + 5.75 + 6.75)/4 = 6.6425$$

The social class level means are obtained by adding up the 6 means for each social class across grades and reinforcement conditions. Thus, for middle social class we have:

$$(5.66 + 6.64 + 6.58 + 5.75 + 8.25 + 9.66)/6 = 7.09$$

All the sample level means are given below:

Grade Level Means: Grade 2: 6.325, Grade 6: 7.29

Social Class Means: Middle: 7.09, Lower: 6.525

Reinforcement Level Means: Mat.: 6.643, Praise: 6.825, Sym: 6.955

The reinforcement means are very close, making it quite likely that they are estimating equal population values, which is reflected in the very small $F = .111$.

Now, let us turn to the 2 way interaction effects that were significant, that is, *AC* (Social class × reinforcement) and *BC* (grade × reinforcement). To interpret these we need the means for social class by reinforcement combined over grade and the means for grade by reinforcement combined over social class. These means are presented below:

		Reinforcement					Reinforcement		
		Mat.	Praise	Symb.			Mat.	Praise	Symb.
Middle		5.71	7.45	8.12	Grade 2		7.04	6.03	5.92
Lower		7.58	6.21	5.79	Grade 6		6.25	7.63	8.00

The mean of 5.71 for the middle class and material reinforcement cell is obtained by adding the means for this set of conditions for the two grades, that is, $(5.66 + 5.75)/2 = 5.71$, and similarly for the other means. The mean of 7.04 for the grade 2 by material reinforcement condition is obtained by adding the means for this set of conditions for the two social classes, i.e., $(5.66 + 8.41)/2 = 7.04$ and similarly for the other means.

The interaction for social class \times reinforcement is disordinal. That is, the lower class children respond better to the material reinforcement and then the means “flip flop”; the middle class children respond better to the praise and symbolic reinforcement.

The grade by reinforcement interaction is also *disordinal*; that is, the younger children respond better to the material reinforcement, and then the means “flip flop”; the older children respond better to praise and symbolic reinforcement.

The Three Way Interaction Effect

Why was the three way interaction not significant? As was mentioned earlier, *a significant three way interaction implies that the two way interaction profiles are different for different levels of the third factor*. If the patterns of means (profiles) are similar, then no interaction will be found. We present the means again below:

		Grade 2			Grade 6		
		Mat.	Praise	Symb.	Mat.	Praise	Symb.
Middle		5.66	6.64	6.58	5.75	8.25	9.66
Lower		8.41	5.41	5.25	6.75	7.00	6.33

Note that the profile of means for second graders is very similar to that for sixth graders. In both cases the mean is higher for lower social class under material reinforcement, and then reverses and is higher for praise and symbolic reinforcement for both grade levels. That is, for both the second and sixth grade we have the same type disordinal interaction.

Now we consider two hypothetical situations in which a significant three way interaction is present, and illustrate these graphically. The first example is a sex \times treatment by race design while the second example involves a counseling methods \times counselors \times sex design. Suppose that the means for the two way design (col-

lapsed on race) and for counseling methods \times counselors (collapsed on sex) were as follows:

<i>Example 1</i>			<i>Example 2</i>	
	T_1	T_2		
Males	60	50	Method 1	85
Females	40	42	Method 2	75
				76

Example 1 shows a clear ordinal interaction while example 2 shows a disordinal interaction, but neither of these tells the whole story. We now present the two way profiles of means for whites and blacks for example 1 and for males and females for example 2:

	Whites		Blacks			Males		Females	
	T_1	T_2	T_1	T_2		C_1	C_2	C_1	C_2
Male	65	50	55	50	Method 1	80	70	90	75
Female	40	47	40	37	Method 2	70	78	80	74

For example 1 we can see that the profiles of means for whites and blacks are distinctly different. Race further moderates the sex by treatment interaction. For whites we have a strong ordinal interaction while for blacks there is no interaction effect. For the counseling example (example 2), we see that the disordinal interaction effect apparent when males and females were combined together was due to the males. We see a clear disordinal method by counselor interaction for males, while for females there is an ordinal interaction. Here, sex moderates the method by counselor interaction. The practical implications for this example are that counselor 1 does uniformly better with method 1 regardless of sex. Method 2 is optimal for counselor 2 working with males, but for females counselor 2 is equally effective with both methods. We display graphically the interaction profiles for these two examples in Figure 4.2. It is important to emphasize again that a significant three way interaction means that the two way interaction profiles for *any two* of the factors are different for the levels of the third factor. Thus, in the sex by treatment by race example above, a significant three way interaction means that:

1. The sex by treatment profiles are different for the races (which is what we illustrated).
2. The sex by race profiles are different for the treatments.
3. The treatment by race profiles are different for the two sexes.

In the context of aptitude-treatment interaction (ATI) research, Cronbach (1975) had an interesting way of characterizing higher order interactions:

When ATI's are present, a general statement about a treatment effect is misleading because the effect will come or go depending on the kind of person treated.... An ATI result can be taken as a general conclusion only if it is not in turn moderated by further variables. If Aptitude \times Treatment \times Sex interact, for example, then the Aptitude \times Treatment effect does not tell the story. Once we attend to interactions, we enter a hall of mirrors that extends to infinity, (p. 119)

Interpreting Patterns of Significant Effects

To further continue our discussion of interpreting effects from a three way ANOVA, we consider an a (methods) \times B (teacher) \times C (sex) example. We examine three possible patterns of significant results, and how one would interpret those patterns.

Pattern 1

A*(method)
B(teacher)
C(sex)
AB
AC
BC
ABC
 $*p < .05$

Here only the method main effect is significant. Since there are no significant interactions we needn't qualify our statement on the efficacy of methods. It can be stated that one method produces *uniformly* higher achievement than the other regardless of teacher and regardless of sex of the child. A pattern of means that would be congruent with the above results is

	T_1	T_2
M_1	72	70
M_2	63	62

Pattern 2

A*(method)
B(teacher)
C(sex)
AB*

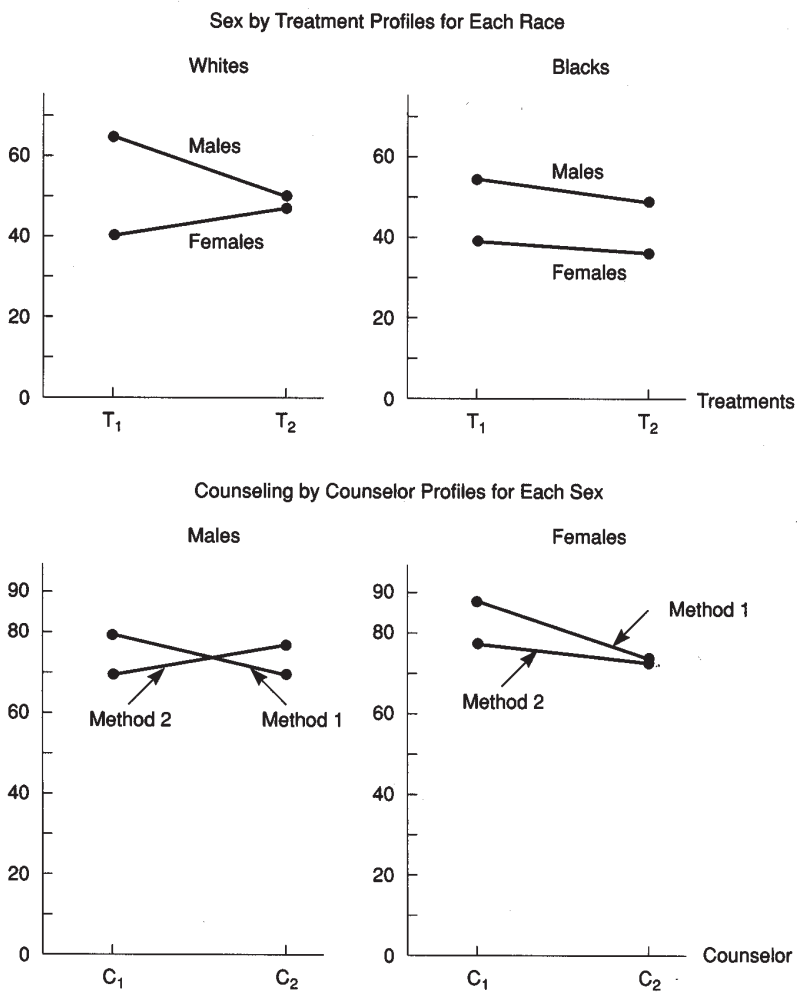


FIGURE 4.2 Two Way Interaction Profiles for Sex by Treatment by Race Design and for the Counseling by Counselor by Sex Design

AC
BC
ABC
 $*p < .05$

Here we have a method main effect again, but this time there is also a significant method by teacher interaction. Thus the efficacy of method needs to be *qualified*. The interaction is telling us that how much better one method is than another depends on the teacher. A pattern of means congruent for the above would be

	T_1	T_2	
M_1	70	65	67.5
M_2	60	62	61.0
	65	63.5	

Method 1 is superior to method 2; however, the degree of superiority depends sharply on the teacher. For teacher 1 method 1 is vastly superior, while for teacher 2 method 1 is only slightly better than method 2.

Pattern 3

A(method)
B*(teacher)
C(sex)
AB
AC
BC
ABC*
 $*p < .05$

The teacher main effect here needs to be considerably qualified because of the significant three way interaction. This could be discussed in terms of differences between two way profiles, as was done previously. Or we might think of it as follows. Call the two teachers Ms. Jones and Mr. Morton. The main effect is telling us that one teacher tends to get higher achievement regardless of method and sex of child (suppose this is Ms. Jones). The three way interaction is telling us that how much better achievement Ms. Jones obtains *depends on both* the method being taught and on the sex of the child. For example, Ms. Jones might do much better

than Mr. Morton working with method 1 and girls, while she gets only slightly better achievement working with method 2 and boys.

Three Way ANOVA on SAS and SPSS

To illustrate the setup of the control lines for running a three way ANOVA on SAS we consider the following sex \times age \times treatment data set:

		Treatments		
		1	2	3
Males	Age			
	14	1 4,6,9 (111)	2 2,3,8 (112)	11,9,16 (113)
	17	9,11,8 (121)	11,7,8 (122)	10,14,9 (123)
Females	14	10,2,8 (211)	12,7,15 (212)	3,7,4 (213)
	17	7,6,12 (221)	9,11,7 (222)	10,15,8 (223)

The numbers in parentheses are the cell identifications, and are very important in identifying to the packages where the data originate. The 111 cell ID means the first level for each factor, while 113 means the first level for factors 1 and 2 and the third level for factor 3, and 213 means the subject is in the second level for factor 1 (female), the first level for factor 2 (age 14) and the third level for factor 3 (treatment 3). Once the cell identification is clear the rest of the setup is relatively straightforward (see Table 4.5).

Selected printout from SPSS GLM for Windows 12.0 for this data is given in Table 4.6. The SPSS options screen (used for obtaining marginal means) is given in Table 4.7.

Calculation of Sums of Squares in Three Way ANOVA

We illustrate here, using definitional type formulas, how some of the sums of squares given in Table 4.6 are obtained, and leave the calculation of the others as exercises. In doing this we link the process to what was done for two way ANOVA, since it is similar. Recall that earlier in calculating the sum of squares for the main effects for A and B the definitional formulas were

$$SS_A = nJ \sum (\bar{x}_{i.} - \bar{x})^2 \quad \text{and} \quad SS_B = nI \sum (\bar{x}_{.j} - \bar{x})^2$$

That is, the row and column means were deviated about the grand mean, and the weighting factor in each case is the number of observations on which each row

TABLE 4.5
SAS Control Lines for Sex \times Age(2) \times Treat (3) ANOVA

```

TITLE 'THREE WAY ANOVA';
DATA THREEWAY;
① INPUT SEX AGE TREAT Y @@;
   LINES;
② 1 1 1 4 1 1 1 6 1 1 1 9
   1 1 2 2 1 1 2 3 1 1 2 8
   1 1 3 11 1 1 3 9 1 1 3 16
   1 2 1 9 1 2 1 11 1 2 1 8
   1 2 2 11 1 2 2 7 1 2 2 8
   1 2 3 10 1 2 3 14 1 2 3 9
   2 1 1 10 2 1 1 2 2 1 1 8
   2 1 2 12 2 1 2 7 2 1 2 15
   2 1 3 3 2 1 3 7 2 1 3 4
   2 2 1 7 2 2 1 6 2 2 1 12
   2 2 2 9 2 2 2 11 2 2 2 7
   2 2 3 10 2 2 3 15 2 2 3 8
PROC PRINT;
PROC GLM;
③ CLASS SEX AGE TREAT;
  MEANS SEX AGE TREAT SEX*AGE
④ SEX*TREAT AGE*TREAT
  SEX*AGE*TREAT;
⑤ MODEL Y = SEX|AGE|TREAT;

```

① In the INPUT statement we list the variables in the analysis.

② The first 3 numbers for each block of 4 numbers is the cell ID, with the fourth number being the score on the dependent variable. Thus, the first subject in cell 111 has a score of 4, the second subject in cell 111 has a score of 6 and the third subject a score of 9. Although not necessary, we have put the data for each cell on a separate line for ease of reading.

③ This CLASS statement lists the grouping variables (factors) for the ANOVA.

④ The MEANS statement is needed to obtain the level means (SEX AGE TREAT), the means for the two way interactions (SEX*AGE SEX*TREAT AGE*TREAT) and the cell means.

⑤ This is the abbreviated way of representing a full three way factorial model in SAS.

TABLE 4.6
SPSS GLM Printout for Three Way ANOVA: Tests of Significance
and Marginal Means for All Effects

Insert art p. 176 of previous edition

TABLE 4.6 (Continued)

Insert art p. 177 of previous edition

TABLE 4.6 (Continued)

Insert art p. 178 of previous edition	
---------------------------------------	--

mean (nJ) or column mean (nl) is based. In calculating the sums of squares for the main effects in three way ANOVA we simply deviate the *level* means about the grand mean, and the weighting factor in each case is the number of observations on which each level mean is based. Now the grand mean for the ANOVA in Table 4.6 is the average of the level means for sex and is thus 8.5555. Since each sex level mean is based on 18 observations, the sum of squares for the sex main effect is:

$$\begin{aligned} SS_{\text{sex}} &= 18[(8.6111 - 8.5555)^2 + (8.5000 - 8.5555)^2] \\ &= .11129 \end{aligned}$$

The discrepancy from the value in Table 4.6 is due to rounding error. Similarly, the sum of squares for treatment is given by

$$\begin{aligned} SS_{\text{tr}} &= 12[(7.6666 - 8.5555)^2 + (8.3333 - 8.5555)^2 + (9.6666 - 8.5555)^2] \\ &= 24.88864 \end{aligned}$$

The calculations for a two way interaction effect are exactly the same as for the two way ANOVA (see Table 4.6), *after* one has collapsed on the levels of the third factor. To illustrate we consider the calculation of sum of squares for the sex by treatment interaction. The means for sex by treatment combined over the two age groups, along with the row and column means and the interaction effects (in parentheses) are:

		Treatments			
		1	2	3	
Sex	1	7.8333 (.1111)	6.5000 (-1.89)	11.5000 (1.78)	8.6111
	2	7.5000 (-.1111)	10.1667 (1.89)	7.8333 (-1.78)	8.5000
		7.6667	8.3333	9.6666	8.5555

Recall that the formula for sum of squares interaction is

$$SS = n \sum \hat{\phi}_{ij}^2$$

where n is the number of observations in each cell and $\hat{\phi}_{ij}$ is the estimated interaction effect for the ij th cell, and

$$\hat{\phi}_{ij} = \bar{x}_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}$$

Thus, the sum of squares here is

$$SS = 6[(.1111)^2 + (-1.89)^2 + (1.78)^2 + (-.1111)^2 + (1.89)^2 + (-1.78)^2] \\ = 80.89$$

The calculations for interaction sum of squares for sex by age and age by treatment are similar, and are left as exercises.

To calculate sum of squares for the three way interaction effect we first compute variability of the cell means about the grand mean (denote this by $SS_{\text{cell(ABC)}}$). The means are:

		Males			Females		
		1	2	3	1	2	3
Age	14	6.3333	4.3333	12.0	6.6667	11.3333	4.6667
	17	9.3333	8.6667	11.0	8.3333	9.0000	11.0000

TABLE 4.7
SPSS 12.0 GLM Options Screen for Obtaining
Marginal Means and Interaction Means

Insert art from p. 179 of previous edition

$$SS_{\text{cell}(ABC)} = 3[(6.3333 - 8.5555)^2 + (4.3333 - 8.5555)^2 + (12 - 8.5555)^2 \\ + \cdots + (9 - 8.5555)^2 + (11 - 8.5555)^2] = 221.556$$

From this quantity we subtract all variation due to the main effects and the first order interactions. What remains is variability due to the three way interaction effect;

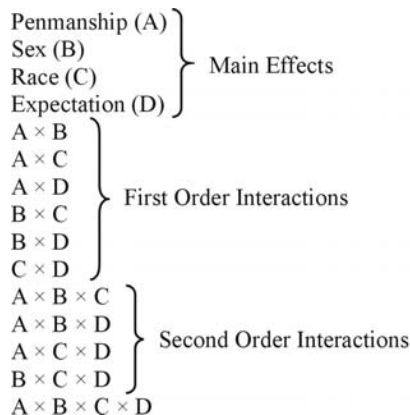
$$SS_{ABC} = SS_{\text{cell}(ABC)} - SS_A - SS_B - SS_C - SS_{AB} - SS_{AC} - SS_{BC} \\ SS_{ABC} = 221.556 - .1111 - 36 - .1111 - 24.8888 - 80.8888 - 4.6667 \\ = 74.889$$

The error term for equal cell size, as was true for two way ANOVA, is simply the *average* of the cell variances.

$$MS_w = (6.3333 + 10.3333 + 13 + \cdots + 10.3333 + 4 + 13)/12 \\ = 9.0555$$

Four Way Analysis of Variance

In a four way analysis of variance we are examining the effect of 4 independent variables on some dependent variable. We consider an example from the literature to illustrate. Chase (1986) examined the effect of penmanship quality, sex, race, and reader expectation on the grade given an essay test. The graders were 80 elementary and middle school inservice teachers in an integrated large urban area of the Midwest. Each grader was given contrived student school records, some of which contained mainly As and Bs while others contained mostly Ds and Us. These records were intended to create in the essay reader a set as to the level of achievement expected from the students whose paper was being graded. Thus there were two levels for reader expectation. The essays were written in poor and good quality of penmanship, as judged by the Ayres handwriting scale. Thus, a $2 \times 2 \times 2 \times 2$ four way ANOVA was run. As mentioned earlier, in a 4 way design there are 16 sources of variation and 15 hypotheses that are tested. For the Chase example the 15 effects are:



One does not see very many 4 or 5 way ANOVAs in the literature. A couple of reasons for this are (1) the difficulty of interpreting higher order interactions and (2) sample size required so that some of cell frequencies are *not* extremely small (like 1 or 2 subjects).

We wish to discuss a caution in using such designs for another reason. While the use of complex ANOVA designs is the only way to get at higher order interactions,

and their “real” existence may have important practical implications, the key word here is real. Remember in a 4 way ANOVA we are testing 15 hypotheses. To some researchers this may seem like a boon, but it can be a bane if one is not careful. Researchers using such designs often interpret any effects that are significant at the .05 level. The potential danger with this for 3, 4, or 5 way ANOVAs is that the overall α level gets out of control. Recall again from Chapter 2 that if we are testing k hypotheses, each at the .05 level, then an upper bound on overall α is given by $1 - (1 - .05)^k$. Below we list the upper bound on overall α for 3, 4, and 5 way ANOVAs if .05 level is used for each effect:

	Three Way	Four Way	Five Way
Number of hypotheses being tested	7	15	31
Upper Bound on Overall α	.30	.536	.79

The results of the Chase study mentioned previously provide a perfect illustration of the danger. In that study the focus was on looking for interactions, but *no specific interactions were hypothesized a priori to be significant*. However, two significant higher order interactions were found at the .05 level, and these were the only significant results. The cell size was equal so that the overall $\alpha = .536$. But this is saying that the probability of *at least one* false rejection is uncomfortably high. Thus, the two significant results found could very well be type I errors or spurious results. At the very least, the reader should be warned of this possibility.

A simple way of controlling the escalating overall α level for 3 and 4 way designs is to test each effect at a more stringent α level, say $\alpha = .01$. Then we are assured that overall $\alpha \leq .07$ for the 3 way and overall $\alpha \leq .15$ for the 4 way design because of the Bonferroni inequality. But the price of this is even worse power for detecting interactions. Now, if sample size is large enough, then we can have the luxury of setting $\alpha = .01$ and still have adequate power. For example, if $n = 270$ in a $2 \times 3 \times 3$ design, then power will be good to adequate for detecting all effects (although marginally so for the BC and ABC interactions).

An Improved Bonferroni Type Procedure

Holland and Copenhaver (1988) discuss several new and improved competitors to the Bonferroni procedure. The one we consider here and illustrate is due to Holm (1979). For this procedure one needs the p values (tail probabilities) for each hypothesis being tested, but since these are printed out on the major statistical packages this is no problem.

The general problem then is to keep overall α under control when testing a set of k hypotheses. The k hypotheses could be of a variety of forms: (1) the numerous F tests from a complex factorial design, (2) the numerous t tests involved if one com-

compares two groups on large set of dependent variables, (3) examining a large number of 2×2 contingency tables from say an original 5×7 contingency table, and (4) determining which of 50 individual between correlations are significant, in analyzing the association between two sets of variables (5 in one set and 10 in the other set).

In the Holm procedure the p values for the k hypotheses are ordered from smallest to largest: $p_{(1)} < p_{(2)} < \dots < p_{(k)}$. Tied p values can be ordered arbitrarily.

Let $H_{(1)}, \dots, H_{(k)}$ denote the hypotheses corresponding to these ordered p values. Suppose i^* is the smallest integer from 1 to k such that

$$p_{(i)} > \alpha/(k - i^* + 1)$$

Then the Holm procedure rejects $H_{(1)}, \dots, H_{(i-1)}$ and retains $H_{(i)}, \dots, H_{(k)}$. The increased power for the Holm procedure comes from the fact that $\alpha/(k - i + 1)$ is larger than α/k .

To illustrate the use of the Holm procedure consider a hypothetical 4 way ANOVA. Suppose we wish to control overall α at .10 for the $k = 15$ hypotheses and that the ordered p values are:

i	$p(i)$	$\alpha/(k-i+1)$
1	.0001	.0067
2	.0001	.0071
3	.0017	.0077
4	.0046	.0083
5	.0053	.0091
6	.0078	.0100
7	.0094	.0111
8	.0113	.0125
9	.0435	.0143
10	.0896	.0167
11	.1342	.0200
12	.2689	.0250
13	.4625	.0333
14	.5813	.0500
15	.6437	.1000

We see here that i^* is 9, where $p(i) = .0435 > .0143$. Thus, by the Holm procedure we would declare 8 effects in the design significant, with assurance that overall $\alpha = .10$. If Bonferroni had been used, an effect would need a $p < .067$ to be declared significant, and only 5 effects would have been significant.

4.5 A COMPREHENSIVE COMPUTER EXAMPLE USING REAL DATA

To tie together several elements discussed in this chapter, we consider a computer analysis of the CARTOON data set. In this study an instructional slide presentation (18 slides) was developed, with the topic being the behavior of people in a group situation, and in particular the various roles or character types that group members often assume. Each role was identified by an animal. Each animal was shown on two slides, once in a cartoon sketch and once in a realistic picture. A random half of the 179 subjects saw the slides in black and white and the other half saw the slides in color. The subjects were immediately posttested for the number of cartoon characters they could identify (CARTOON 1) and for the number of realistic characters they could identify (REAL1). They were retested 4 weeks later on the same two variables. Three groups of subjects were involved in the study: preprofessional and professional personnel from three hospitals and a group of Penn State college students. For the computer analysis, in contrast to the first edition of this text, I just consider the REAL2 variable for this 2×3 design (color of presentation by type of subject). Our analysis serves to illustrate and integrate several aspects of practical data analysis.

The analysis was run on SPSS MANOVA, and only on subjects for which there is complete data. Note that this reduces the effective sample size substantially (from 179 to 105). Immediately we encounter a couple of problems that typify “real world” data analysis. First, the cell sizes are sharply unequal. Second, there is a fair amount of missing data (many subjects did not show up for the retest). Missing data is a fairly common occurrence in certain areas of research, and there is no simple solution for this problem. If it can be assumed that the data is missing *at random*, then there is a sophisticated procedure available for obtaining good estimates (Johnson & Wichern, 1988, pp. 197–202). On the other hand, if the random missing data assumption is not tenable (usually the case), then there is no general consensus as to what should be done. There are various suggestions, like using the mean of the scores on the variable as an estimate, or using regression analysis (Frane, 1976). Probably the “best” solution is to make every attempt to minimize the problem before and during the study, rather than having to manufacture data. The statistical packages SAS and SPSS have various ways of handling missing data. The default option for both, however, is to delete the case if there is missing data on any variable for the subject (called listwise deletion).

Now recall that the homogeneity of variance assumption for factorial designs is that the *cell* population variances are equal. Since the cell sizes are sharply unequal, a violation of this assumption will distort the type I error rate, and it is important to check this assumption. Fortunately this assumption is tenable (using Cochran’s test, $p = .401$).

TABLE 4.8
SPSS MANOVA Control Lines and Selected Printout for CARTOON Data

```

TITLE 'TWO WAY ANOVA ON CARTOON DATA FOR REAL2'.
DATA LIST FIXED/ID 1-3 COLOR 5 ED 7 LOCATION 9 OTIS 11-13 CARTOON1
  15 REAL1 17 CARTOON2 19 REAL2 21.
BEGIN DATA.

DATA (FROM 3.5 FLOPPY DISK)

END DATA.
MANOVA REAL2 BY COLOR(0,1) ED(0,2)/
  PRINT = CELLINFO(MEANS)/.

Tests of Significance for REAL2 using UNIQUE sums of squares
Source of Variation      SS      DF      MS      F      Sig of F

WITHIN CELLS              532.74    99      5.38
COLOR                     15.04     1     15.04    2.79    .098
ED                        79.95     2     39.98    7.43    .001
COLOR BY ED               12.62     2      6.31    1.17    .314

(Model)                   95.89     5     19.18    3.56    .005
(Total)                   628.63   104     6.04

R-Squared =                .153
Adjusted R-Squared =      .110

```

The significant main effect for REAL2 is an overall test that merely tells us the three population column means differ. It does not indicate which particular column means are different. For this we need a post hoc procedure, just as we did in one way ANOVA. We use the Tukey procedure. Recall that for one way ANOVA the endpoints for the confidence intervals were given by

$$(\bar{x}_i - \bar{x}_j) \pm q_{\alpha; k, N-k} \sqrt{MS_w / n}$$

where n was the assumed common group size. Remember also that when the group sizes were unequal the Tukey was still applicable, provided that the population variances were equal and that n was replaced by the harmonic mean for each pair of groups.

In application of the Tukey to factorial designs the n is replaced by the number of observations on which each row or column mean is based, for equal cell size. When the cell sizes are unequal, as in the study we are examining, we again employ the harmonic mean, but now for each pair of row and/or column sizes.

Below we present the cell sizes for the REAL2 variable. From this table we see that the column sizes are 24, 26, and 55.

<i>Real2</i>			
<i>Preprof</i>	<i>Prof</i>	<i>Coll</i>	<i>Row Mean</i>
3.667	3.937	4.852	4.33
(2.43)	(2.05)	(2.23)	
<i>n</i> = 12	<i>n</i> = 16	<i>n</i> = 27	
2.333	2.700	4.964	3.88
(1.97)	(1.89)	(2.73)	
<i>n</i> = 12	<i>n</i> = 10	<i>n</i> = 28	
3.0	3.46	4.91	4.11

Thus, the harmonic means for each pair of column sizes are given by

$$2(24)(26)/50 = 25, 2(24)(55)/79 = 33.42 \text{ and } 2(26)(55)/81 = 35.31$$

Below are the calculations and the intervals:

Groups Compared	Harmonic Mean	Critical Value	Interval
Prof-Preprof Mean diff = .46	25	$3.356\sqrt{5.381 / 25} = 1.56$	(-1.1, 2.02)
Coll-Preprof Mean diff = 1.91	33.42	$3.356\sqrt{5.381 / 33.42} = 1.35$	(.56, 3.26)
Coil-Prof Mean diff = 1.45	35.31	$3.356\sqrt{5.381 / 35.31} = 1.31$	(.14, 2.76)

These intervals show that the college students differed significantly from both the preprofessional and the professional groups, and examination of the column means in the above table shows that the college students scored higher in each case.

4.6 POWER ANALYSIS

Power Estimation for Two Way Analysis of Variance

Because we are basing our treatment of power on Cohen's book (1977, revised edition), it is very important to note here that estimation of power for factorial designs changed significantly from the first edition (Cohen, 1969) to the second edition. We quote from a couple of footnotes in Cohen's (1977) edition:

Readers familiar with the first edition should note that the treatment of main effects (and even more so, of interactions) in factorial design differs considerably here. The

systematic overestimation of power for main effects by the former method proved to be unacceptably large in some applications. The present method gives quite accurate and unbiased results.... In the case of interactions, both the *ES* (effect size measure) of formula (8.3.6) below and the *n* used for table entry have been changed in this edition, thus avoiding substantial underestimation of power for interactions, (footnotes 2 and 3, p. 364 and p. 369)

The main reason for setting up a factorial design is to test for an interaction effect. Unfortunately, as we will see shortly, the power for detecting this interaction can be inadequate. Suppose that we have an $A \times B$ design, with r levels for factor A and c levels for factor B . The effect sizes for the main effects and interaction may be expressed as follows:

$$A \text{ main effect: } \hat{f}_A \sqrt{(r-1)F_A / N} \quad (5)$$

$$B \text{ main effect: } \hat{f}_B \sqrt{(c-1)F_B / N} \quad (6)$$

$$AB \text{ interaction: } \hat{f}_{AB} \sqrt{(r-1)(c-1)F_{AB} / N} \quad (7)$$

To illustrate the power calculations we consider a 2×3 design with 10 observations per cell:

		<i>B</i>		
	10	10	10	30
<i>A</i>				
	10	10	10	30
	20	20	20	60 = <i>N</i>

It might appear that the n we should use to enter the power tables for the A main effect is 30, since the row mean is based on 30 subjects. However, a slight adjustment is necessary (Cohen, 1977, p. 365). The same is true for the B main effect and the interaction, and the following n s are what are needed:

<i>Effect</i>	<i>n used to enter the table</i>
<i>A</i> main effect	$n_A = [N - rc]/r + 1$
<i>B</i> main effect	$n_B = [N - rc]/c + 1$
<i>AB</i> main effect	$n_{AB} = [(N - rc)/((r-1)(c-1) + 1)] + 1$

The F ratios from the above study, along with the corresponding effect sizes and power are presented below:

<i>Effect</i>	<i>F</i>	<i>Effect Size</i>	<i>n to enter table</i>	<i>Power</i>
<i>A</i> main	1.8	$\sqrt{1(1.8) / 60} = .173$	$(60 - 6) / 2 + 1 = 28$.25
<i>B</i> main	3.6	$\sqrt{2(3.6) / 60} = .346$	$(60 - 6) / 3 + 1 = 19$.64
<i>AB</i> inter	2.1	$\sqrt{2(2.1) / 60} = .265$	$[(60 - 6) / 2 + 1] + 1 = 19$.40

Since there is a fairly large effect size for the B main effect, power is at least fair. However, for the interaction with a medium effect size, power is poor.

Power Estimation for Three Way Analysis of Variance

Power analysis for three way ANOVA is a straightforward generalization of that for two way ANOVA. Consider a general three way design: factor A — a levels, factor B — b levels, and factor C — c levels. We simply need the following effect size for the three way interaction:

$$\hat{f} = \hat{\sigma}_{ABC} / \hat{\sigma}, \text{ where } \hat{\sigma}^2 = MS_w \text{ and } \hat{\sigma}_{ABC} = SS_{ABC} / N$$

It can be shown that \hat{f} is related to the F statistic for the three way interaction as follows:

$$\hat{f} = \sqrt{[(a-1)(b-1)(c-1) / N] F_{ABC}} \quad (8)$$

In the above, $u = (a-1)(b-1)(c-1)$ is the degrees of freedom for the three way interaction. The n that is used to enter the power tables for *each* effect in the design (i.e., main effects, first order interactions and the second order interaction) is

$$n' = (N - abc) / (u + 1) + 1$$

where $(N - abc)$ is the degrees of freedom for the error term, and u is the degrees of freedom for the effect in question: For example, $(a-1)$ for the A main effect, $(b-1)$ for the B main effect, $(a-1)(b-1)$ for the AB interaction, etc.

To illustrate power estimation we consider a $2 \times 3 \times 3$ design with 5 subjects per cell. For example, the design might be sex by treatments by social class. The power values are given for differing α and for different effect sizes.

Power as a Function of Effect Size & α Level
in a $2 \times 3 \times 3$ Design with $n = 5$

Effect	u	n'	$f = .10$		$f = .25$		$f = .40$	
			.05	.10	.05	.10	.05	.10
<i>A</i>	1	37	.13	.22	.58	.70	.93	.96
<i>B</i>	2	25	.10	.19	.47	.60	.87	.93
<i>C</i>	2	25	.10	.19	.47	.60	.87	.93
<i>AB</i>	2	25	.10	.19	.47	.60	.87	.93
<i>AC</i>	2	25	.10	.19	.47	.60	.87	.93
<i>BC</i>	4	16	.09	.16	.38	.51	.81	.88
<i>ABC</i>	4	16	.09	.16	.38	.51	.81	.88

The power values at $\alpha = .05$ for small and medium effect sizes are boxed in. Notice that almost all (only 1 exception) these values are less than .50, that is, poor.

In Chapter 3 on power analysis we indicated that SPSS MANOVA can be used to obtain estimates of power for various fixed effects univariate and multivariate tests, and showed in Table 3.2 the control lines for obtaining the power estimates for the t test and a one way ANOVA. To obtain power estimates for the various effects in a factorial design is equally as simple. For example, for the data on page 174, we simply insert after the MANOVA command the following subcommands

```
POWER = F(.05) /  
PRINT = CELLINFO(MEANS) SIGNIF(EFSIZE) /
```

Directly beneath the ANOVA table SPSS prints out the effect sizes and power estimates for each effect in the design. It looks like this:

Placeholder for art p. 192 of previous edition

Recall from Chapter 3 that a partial η^2 around .01 indicates a small effect size, a partial η^2 around .06 a medium effect size, and a partial η^2 around .14 was a large effect size. For the above design there are three large effect sizes (for age main effect, the sex by treat interaction, and for the three way interaction). Recall that the above two interaction effects were significant at the .05 level, while the age main effect was not significant. Here, because of the very small sample size (only 3 subjects per cell), power was only adequate (around .70) when the effect size was very large.

4.7 FIXED AND RANDOM FACTORS

At this point it is important to distinguish between fixed and random factors. All that we have considered to this point is what is called *fixed effects* ANOVA. For example, in comparing three different diets (one way ANOVA), the diets are not randomly sampled from some population of diets, but rather they are fixed by the experimenter. Furthermore, the experimenter is not interested in generalizing to some population of diets but wishes to determine which of the diets in the study is superior to the others. Thus, inferences in the study are “fixed” or limited to the diets under consideration. There are situations in factorial designs where the experimenter may wish to generalize beyond the given levels of a factor in the study, and in this case the factor is considered random. Let us consider two examples to illustrate.

First, suppose we want to compare three different teaching methods (fixed factor) in 7 randomly selected schools in some metropolitan area. The investigator wishes to generalize to the population of schools in this area. Schools is the random factor and we have what is called in the literature a *mixed model*, since one factor (methods) is fixed while the other factor is random.

As a second example suppose we are comparing the effect of two reading methods on comprehension for second graders. We select 5 stories that we consider to be representative of second grade reading material. We have a 2(methods) \times 5(stories) design. We wish to generalize the results to all stories, so that stories is the random factor.

A random factor(s) in the design introduces another complication; different error terms (something other than MS_w) are needed for testing some of the effects for significance. For instance, for the teaching methods by schools example, while MS_w is appropriate for testing the school main effect and the interaction effect for significance, the method \times school interaction mean square is the appropriate error term for testing the method main effect.

4.8 SUMMARY

1. In two way ANOVA we are examining the effect of two independent variables (factors) on some dependent variable.

2. For an $A \times B$ design there are 3 hypotheses to be tested: The A main effect (that the population row means are equal), the B main effect (that the population column means are equal), and the $A \times B$ interaction effect.

3. An interaction means that the effect one factor has on the dependent variable is not the same for all levels of the other factor. Two types of interaction, ordinal and disordinal, were discussed.

4. The same error term, MS_w , is used for testing each of the 3 effects. It is a pooled estimate of within cell variability, and for equal cell size is just the average of the cell variances.

5. For balanced designs (equal cell n) the sums of squares are independent, although the F tests are not independent because they share a common error term. However, for total N even moderately large the amount of dependence is small and can be ignored for practical purposes.

6. *For disproportional cell size the sums of squares are correlated (confounded).* Several methods have been suggested in the literature for analyzing such designs. There is now a consensus that generally the regression approach, where the unique contribution of each effect is obtained, should be used. If, however, an a priori ordering of the effects can be established, then the sequential sum of squares approach makes sense.

7. The regression approach, which yields the unique variation due to each effect, is denoted by type III sum of squares in SAS and SPSS for Windows.

8. Aptitude \times treatment interaction (ATI) is a broad area of research that uses factorial designs, and is concerned with the possible moderating effect any individual difference characteristic (sex, age, locus of control, etc.) of subjects may have on their response to treatments.

9. In a three way ANOVA there are 7 hypotheses that are tested. The 3 main effects test whether the population level means are equal. The nature of the two way interactions is ascertained by examining the means for each pair of factors lumped over the third factor. The three way interaction indicates whether the patterns of means (profiles) for any two factors are different for the levels of the third factor.

10. Power estimation for two and three way ANOVA was discussed using Cohen's approach.

11. An improved Bonferroni type procedure, which makes use of p values, is discussed and illustrated.

12. A comprehensive computer example, using real data, is used to illustrate and integrate several important concepts from the chapter, as well as indicating some aspects of practical data analysis.

13. For 3 and 4 way ANOVA there are many hypotheses being tested (7 for three way and 15 for four way). It is important to note that if the .05 level is used for each effect, then the overall α level becomes quite high. Thus, 1 or 2 significant results from such a design, if not hypothesized a priori, could well be spurious.

14. The distinction between fixed and random factors is illustrated with some examples.

EXERCISES

1. Can you think of a fourth advantage of a factorial design?
2. Consider the following hypothetical data for an Age by Treatments factorial design:

		TREATMENTS		
		1	2	3
AGE	10 yrs	21, 27, 23	24, 32, 30	19, 30, 27
		28, 20	35, 32	20, 21
	12 yrs	18, 25, 27	24, 16, 18	34, 28, 21
		20, 23	19, 20	30, 29

- (a) Test each of the effects for significance at the .05 level using the definitional formulas given in the text. Use your calculator to obtain the mean and variance for each cell and then go from there.
 - (b) Which of the effects, if any, are significant? Interpret any effect which is significant.
3. Run problem 2 on the SAS GLM program.
4. Suppose that a study like that for problem 2 had been conducted, starting with 5 subjects per cell, but that for various reasons several subjects dropped out of the study leaving the following disproportional cell size data set:

		TREATMENTS		
		1	2	3
AGE	10 yrs	21, 27, 23	24, 35, 32	19, 30,
		28, 20		20, 21
	12 yrs	25, 20	24, 16, 18	28, 21
			19, 20	30, 29

- (a) Run this data set on both SAS GLM and on SPSSX MANOVA. Which effects are significant at the .05 level? Interpret any significant effects.
- (b) Are the Type I and Type III sums of squares different for all the effects? Explain.
5. Consider the following results from a 2×3 factorial ANOVA (4 subjects per cell) study by Pukulski (*The Reading Teacher*, 1970, 515–522):

Source of Variation	df	MS	F
Sex	1	308.16	.528
Reinforcement	2	3251.29	5.57*
Sex by Reinforcement	2	1094.29	1.87
Error	18	583.78	

* $p < .05$

- (a) Estimate what his power was at $\alpha = .10$ for detecting the reinforcement main effect?
- (b) Estimate power at $\alpha = .10$ for detecting the interaction effect.
- (c) Given the result in (b), what would you recommend Pukulski do in a followup study?
6. Explain what Cronbach meant when he said, “Once we attend to interactions we enter a hall of mirrors that extends to infinity.”
7. Suppose an investigator in a heuristic study has a two way design and 5 dependent variables. He runs 5 univariate two way ANOVAs, that is, he does a two way ANOVA on each dependent variable separately. Four of the effects are significant at the .05 level, and he is excited by these results and discusses them in some detail.
- (a) What is the total number of statistical tests that was done here?
- (b) What is the upper bound on overall α ?
- (c) Given the result in (b), should the investigator be excited, or should he be cautiously optimistic?
8. Consider again the method by teacher by sex example on p. 150.
- (a) Suppose that only the AC interaction was significant. Interpret what this result means. Give a pattern of means that is congruent with the above result.
- (b) Suppose that the A and C main effects and the AC interaction were the only significant effects. Interpret these results, and give a pattern of means that is congruent with these results.

9. In Section 4.4 we indicated, using definitional formulas, how various sums of squares for a three way ANOVA are calculated. Finish the calculations for that example by
 - (a) calculating the age by treatment sum of squares
 - (b) calculating the sex by age sum of squares
 - (c) calculating the age main effect sum of squares
10. Construct the treatment by race profiles for example 1 on p. 149 and interpret.
11. An investigator has a 3×3 (treatments by social class) factorial design and from previous literature anticipates a medium treatment main effect and a medium interaction effect. She wishes to know if having 10 subjects in each cell will yield adequate power ($> .70$) for detecting these two effects. Given these results, what is the estimated power for detecting the interaction at $\alpha = .10$? Is power now adequate? Obtain a somewhat rough estimate (since it involves extrapolation) of power at $\alpha = .15$.
12. Suppose an investigator actually hypothesized a significant three way interaction effect (don't expect to find this very often in the literature). He wishes to detect a medium or larger three way interaction effect with power $= .70$ at $\alpha = .10$. He has a $2 \times 2 \times 3$ design. How many subjects per cell are needed?
13. A study by Tuckman, Steber, and Hyman (1979) had principals rate teachers in their schools, whom they had previously nominated as effective or ineffective, on the four dimensions of the Tuckman Teacher Feedback Form: creativity, dynamism, organized demeanor, and warmth and acceptance. There were 180 teachers rated, one-third each at the elementary, intermediate, and high school levels. The primary question in the study, in the authors words was, "Do principals' judgments across the four dimensions of teaching style vary from elementary to intermediate to senior high school principals? That is, do principals at the three levels perceive the four dimensions differently as elements of effective versus ineffective teaching?" They hypothesized elementary principals would see warmth and acceptance and creativity as contributing most to the discrepancy between most effective and least effective teachers while dynamism and organized demeanor were expected to be higher in importance for intermediate and high school principals.
 To test their hypothesis, four two way ANOVAs were run, a separate ANOVA for each dimension of the TTFF. The independent or grouping

variables are school level and effective-ineffective dimension. The following results were obtained:

SOURCE SCHOOL	CREATIVITY			DYNAMISM		ORGANIZED DEMEANOR		WARMTH AND ACCEPTANCE	
	DF	MS	F	MS	F	MS	F	MS	F

The asterisks indicate those effects with a *p* value less than .01.
Also, the following means were obtained on the four variables for the six cells in the factorial design:

	CREATIVITY		DYNAMISM		ORGANIZED DEMEANOR		WARMTH AND ACCEPTANCE	
	M.E.	L.E.	M.E.	L.E.	M.E.	L.E.	M.E.	L.E.
ELEM.	27.3	22.4	25.7	28.9	34.8	27.9	39.3	23.9
INTERM.	29.2	21.8	27.9	22.8	36.8	27.0	35.6	26.5
SENIOR	24.9	15.9	28.2	17.6	36.3	24.4	31.7	26.7

- (a) What are the significant interaction effects on DYNAMISM and WARMTH AND ACCEPTANCE telling us?
 - (b) Explain why these interaction effects occurred, using the cell means.
 - (c) What part(s) of their hypotheses are confirmed by the above analysis?
 - (d) Is further analysis necessary to validate or invalidate some of their hypotheses?
14. A study by Bryan (1974) investigated the peer popularity of learning disabled children. The learning disabled and a sample of control “normal” subjects each consisted of 35 white and 29 black boys and 10 white and 10 black girls. The children were in grades 3, 4, and 5. A combination of two sociometric techniques was used to assess peer popularity. The measures included: (a) the choice of three classmates as friends, classroom neighbors, and invitees to a birthday party; (b) the choice of three classmates who are not friends or neighbors or invitees to a birthday party; and (c) the Guess Who Technique. Sample items from this procedure include: “Who finds it hard to sit still in class? Who is handsome or pretty? Who is always worried or scared?” the scores of the children on items from the above three categories were the sum of the number of classmates who nominated the subject on that item, divided by the total number of votes cast within the classroom. The relationships among the items indicated that the 20 items could be divided into two scales: social acceptance and social rejection.

These were the two dependent variables for the study. The percentages on these two variables were transformed into arc sine equivalents before analysis. This transformation is appropriate when there is reason to believe there may be a relationship between the means and the variances. Such is the case when the dependent variable involves proportions or percentages (see Myers, 1979, p. 73), as in this study. Subjects were cross classified on group (learning disabled or control), sex, and race, and three way ANOVA's were run on each dependent variable, using a least square analysis. The following results were obtained:

	Social Acceptance		Social Rejection	
	Mean Sq	<i>F</i> <i>df</i> = 1,160	Mean Sq	<i>F</i> <i>df</i> = 1,60
Group (A)	.809	19.896***	.589	9.118**
Sex (B)	.149	3.667	.004	.055
Race (C)	.000	.008	.007	.112
A × B	.032	.797	.313	4.850*
A × C	.233	5.737**	.932	14.415***
B × C	.001	.019	.029	.447
A × B × C	.094	2.320	.173	.104

**p* < .05
 ***p* < .01
 ****p* < .001

- Why is the degrees of freedom for error = 160?
- What is the numerical value of the error term for social acceptance and social rejection?
- The cell sizes in the study were unequal, but the exact cell sizes were not reported. Given this, should the author have checked the homogeneity of variance assumption? Why, or why not?
- The author presents the following table for interpreting the significant AC interactions for social acceptance and rejection:

	SOCIAL REJECTION		SOCIAL ACCEPTANCE	
	WHITE	BLACK	WHITE	BLACK
LEARNING DISABLED	15	8	4	6
CONTROL	5	9	10	7

What type of interaction (ordinal or disordinal) resulted for social rejection? for social acceptance? Does there seem to be a particular cell that is primarily responsible for the interactions in each case?

- Calculate the effect size for the a × B × C interaction on social acceptance.

Is this a practically significant effect which we failed to detect because of inadequate power?

15. Run a three way ANOVA on the data given below for a sex(2) \times age(2) \times treat(3) design.

SEX	AGE	TREAT		
		1	2	3
1	1	19,16,18,17	23,24,25,28	16,12,24,10
	2	20,17,18,19	27,31,28,25	19,18,23,27
2	1	17,18,14,22	26,19,13,17	15,17,15,12
	2	13,18,20,19	14,13,21,18	14,18,19,11

- (a) Test each of the effects at the .01 level. Which are significant?
(b) Interpret any significant effects using the appropriate means.
16. Consider the following subset of data (all of site 1) for a SETTING (1 for home and 2 for school) by VIEWCAT (1 for rarely watches Sesame St to 4 for watches the show on average of more than 5 times a week) factorial design. The dependent variable is LETDIFF = POSTLET – PRELET, that is, a measure of how much the children have gained in their knowledge of letters:

	VIEWCAT1	VIEWCAT2	VIEWCAT3	VIEWCAT4
HOME	0, 4	6, 4, 10, 9	14, 7, 7, 28, 16, 32	27, 21, 4
		11, 1, -2, 6, -10	8, -1, 10, 26, -22	24, 35, 5
SCHOOL	7, 3, 8, -1	-1, 4, 17, 6	11, 32, 10, 33, 14	6, 6, 4, 4
	2, -1, -1	9, 21, 5, 7	33, 30, 31, 5	24, -15, 8, 7

- (a) Run a two way ANOVA on this data using either SAS or SPSS. For both the default is the unique variability due to each effect (called type III sum of squares). Which effect(s) is(are) significant at the .05 level?
(b) Using the appropriate means, interpret each of the significant effects.
17. Consider the following approximate 33% random sample of the ATTITUDE data. Here we focus on the SEX and GRADE factors and the change in mathematics attitude (CHGMATH).

	GRADE			
	3	4	5	6
MALE	-1,0, 2,2 -1,0, 2,2	0, -3, 0, -1 3, -5, 0, 3	-1, 1, 1, -1	1, 2, 0, 3, 1 1, -4, 1, 0
FEMALE	3, -2, -2, 2, -2 -3, -1, 0, 0, -1, 1	-1, -1, -3, 0 2, 0, 0, 0	-1, -1, 0, 0 0, 0, -1, 0, 0	-1, 1, 1, -4, -1 0, 4, 1, -1, 0

- (a) Using either SPSS or SAS, test each of the effects for significance at the .05 level. Which, if any, are significant?
18. Run the HEADACHE data on SPSS or SAS, using UNCOMF as the dependent variable. What effects are significant at the .05 level?
 19. Why is it indicated that with real data one will generally have unequal cell size?
 20. In a 5 way ANOVA, how many sources of variation are there?

APPENDIX DOING A BALANCED TWO WAY ANOVA WITH A CALCULATOR

1. Obtain the mean and variance for each cell.
2. Obtain the row, column, and grand means.
3. Obtain the error term (MS_w) as the average of the cell variances.
4. Obtain the sum of squares and mean squares for the main effects.
5. Test each of the main effects for significance.
6. Obtain the sum of squares and mean square for the interaction effect.
7. Test the interaction effect for significance.

To illustrate the above process, we consider the following age by treatment design:

		Treatments		
		1	2	3
AGE	10 yrs	21, 27, 23 28, 20	24, 32, 30 35, 32	19, 30, 27 20, 21
	12 yrs	18, 25, 27 20, 23	24, 16, 18 19, 20	34, 28, 21 30, 29

The means, cell variance, and the row, column, and grand means are as follows:

		TREATS		ROW MEANS	
AGE	23.8 (12.7)	30.6 (16.8)	23.4 (23.3)	25.93	
	22.6 (13.3)	19.4 (8.8)	28.4 (22.3)	23.47	
COLUMN MEANS		23.2	25	25.9	24.70 (GRAND MEAN)

Now we move to step 3 and obtain the error term: Recall from the chapter that for equal cell size the error term is just the average of the cell variances. Therefore,

$$MS_w = (12.7 + 16.8 + 23.3 + 13.3 + 8.8 + 22.3) / 6 = 16.2$$

In step 4 we obtain the sum of squares and mean squares for the main effects (note that cell size = 5):

$$\begin{aligned}
SS_{age} &= 15[(25.93 - 24.7)^2 + 23.47 - 24.7)^2] = 45.39 \\
MS_{age} &= 45.39 / 1 = 45.39 \\
SS_{trts} &= 10[(23.2 - 24.7)^2 + (25 - 24.7)^2 + (25.9 - 24.7)^2] \\
&= 37.8 \\
MS_{trts} &= 37.8 / 2 = 18.9
\end{aligned}$$

In step 5 we test the main effects for significance (we use the .05 level here).

$$F_{age} = MS_{age} / MS_w = 45.39 / 16.2 = 2.80$$

Since the critical value at the .05 level, based on 1 and 24 degrees of freedom, is 4.26, we fail to reject.

$$F_{trts} = MS_{trts} / MS_w = 18.9 / 16.2 = 1.17$$

Here the critical value, based on 2 and 24 df, is 3.40, and we once again fail to reject.

In step 6 we obtain the sum of squares and mean square for the interaction effect. Recall that the sum of squares for interaction involved cell interaction effects, and that each cell interaction effect, in words, is given by cell interaction = cell mean + grand mean - row mean - column mean. The cell interaction effects are given below:

		TREATMENTS	
	-.63	4.37	-3.74
AGE			
	.63	-4.37	3.74

Therefore,

$$\begin{aligned}
SS_{int} &= 5[(-.63)^2 + (4.37)^2 + (-3.74)^2 + (.63)^2 + (-4.37)^2 + (3.74)^2] \\
&= 334.814 \\
MS_{int} &= 334.814 / 2 = 167.407
\end{aligned}$$

Finally, in step 7 we test the interaction effect for significance:

$$F_{int} = MS_{int} / MS_w = 167.407 / 16.2 = 10.33$$

The critical value is 3.40. Thus, there is a significant interaction effect.

Repeated Measures Analysis

CONTENTS

- 5.1 Introduction
- 5.2 Advantages and Disadvantages of Repeated Measures Designs
- 5.3 Single Group Repeated Measures
- 5.4 Completely Randomized Design
- 5.5 Univariate Repeated Measures Analysis
- 5.6 Assumptions in Repeated Measures Analysis
- 5.7 Should We Use the Univariate or Multivariate Approach?
- 5.8 Computer Analysis on SAS and SPSS for Example
- 5.9 Post Hoc Procedures in Repeated Measures Analysis
- 5.10 One Between and One Within Factor—A Trend Analysis
- 5.11 Post Hoc Procedures for the One Between and One Within Design
- 5.12 One Between and Two Within Factors
- 5.13 Totally Within Designs
- 5.14 Planned Comparisons in Repeated Measures Designs
- 5.15 Summary

5.1 INTRODUCTION

In our discussion of one way ANOVA and factorial ANOVA the subjects were only measured once on the dependent variable. In this chapter we consider designs that measure subjects several times, either on the same dependent variable or on different measures. The simplest repeated measures design measures the subjects twice, with an intervening treatment. Schematically, we have

Pretest Treatment Posttest

In this case the student may recall that the t test for correlated (dependent) samples applies. Repeated measures analysis of variance (where the subjects are mea-

sured more than twice) is the generalization of the t test for correlated samples, just as ANOVA (k groups) was the generalization of the t test (two groups) for independent samples.

There are many situations in which repeated measures are either appropriate or the natural thing to use. For example, if we are concerned with performance trends over time. Bock (1975) presented an example comparing boys' and girls' performance on vocabulary acquisition over grades 8 through 11. Here the focus is often on the mathematical form of the trend, that is, whether it is linear, quadratic, etc. The same type of analysis applies whether we are concerned with cognitive variables (as above), or personality changes for a group of subjects over time, or developmental (physiological) changes for a group of infants (children).

Another class of repeated measures situations occurs when we are comparing the same subjects under several different treatments (drugs, stimulus displays of different complexity, etc.). For example, we may be interested in the effects of 4 drugs on reaction time for a group of subjects, or in the effects of repeated practice (say over 3 sessions) on a learning task.

Another useful application of repeated measures occurs in combination with a one way ANOVA design. In a one way design involving treatments the subjects are posttested to determine which treatment is best. If we are interested in the lasting or residual effects of treatments, then we need to measure the subjects a few more times. Huck, Cormier, and Bounds (1974) present an example in which three teaching methods are being compared, but in addition the subjects are again measured 6 weeks and 12 weeks later to determine the residual effect of the methods on achievement. A repeated measures analysis of such data *could* yield a quite different conclusion as to which method might be preferred. Suppose the pattern of means looked as follows:

	<i>Posttest</i>	<i>Six Weeks</i>	<i>12 Weeks</i>
Method 1	66	64	63
Method 2	69	65	59
Method 3	62	56	52

Just looking at a one way ANOVA on posttest scores (if significant) could lead one to conclude that method 2 is best. Examination of the pattern of achievement over time shows however that for lasting effect method 1 is to be preferred, because after 12 weeks the achievement for method 1 is superior to method 2 (63 vs. 59). What we have here is an example of a method by time interaction effect.

Another class of situations in which repeated measures designs apply is when the same subjects are given a series of tests or subtests. For instance, Glass and Hopkins (1984) present the following example. A group of 12 neurologically handicapped children are measured on the information, vocabulary, digit span, and

block design subtests of the Wechsler Intelligence Scale for Children (WISC). If the 12 fall into, say, 3 different types of neurological handicaps, then we may be interested in whether certain deficits on WISC are particularly associated with different types of handicaps. Here a subject by subtest interaction is the focus.

In this chapter we consider repeated measures designs of varying complexity. The simplest design involves a single group of subjects measured under various treatments (conditions), or at different points in time. Schematically, it looks like this:

		Treatments				
		1	2	3	k
Subjects	1					
	2					
	n					

We then consider a one between and one within design. Many texts use the terms “between” and “within” in referring to repeated measures factors. A between variable is simply a grouping or classification variable such as sex, age, or social class. A within variable is one on which the subjects have been measured repeatedly (like time). Some authors even refer to repeated measures designs as within designs (Keppel, 1983). An example of a one between and one within design is

		DRUGS		
		1	2	3
Schizophrenics Depressives				

Here the *same* schizophrenics and depressives are given three drugs to determine which of them is best in inhibiting some undesirable response. The teaching methods study mentioned previously is another example of a one between and one within design, where methods is the between variable (different subjects taught by different methods) and time is the within variable. The reader should be aware that there are three other names that are used by some authors for the same design: Lindquist Type I, split plot, and two way ANOVA, with repeated measures on one factor.

Next we consider a one between and two within design. As an example, suppose a researcher in child development is interested in observing three groups of children (ages 3, 4, and 5) in two situations at two different times (morning and afternoon) of the day. She is concerned with the extent of their social interaction, and will measure this by having two observers independently rate the amount of social interaction. The average of the two ratings will serve as the dependent variable. The age of the children is the grouping or between variable here. The two within

variables are situation and time of day. There are four scores for each child: social interaction in situation 1 in the morning, social interaction in situation 1 in the afternoon, social interaction in situation 2 in the morning, and social interaction for situation 2 in the afternoon.

Schematically, the design is as follows:

	SITUATION	1		2	
		Morn.	After	Morn.	After
AGE	3 years	y_1	y_2	y_3	y_4
	4 years				
	5 years				

where the y s represent the four social interaction measures for each subject. One can think of this as a three way ANOVA, but it is a *different* type of analysis of variance from that in Chapter 4, because the subjects' scores are correlated across situation and across time, and this must be taken into account in the analysis.

Finally, we discuss planned comparisons in repeated measures designs.

5.2 ADVANTAGES AND DISADVANTAGES OF REPEATED MEASURES DESIGNS

Recall that the two basic objectives in experimental design are elimination of systematic bias and the reduction of error (within gp or cell) variance. The main reason for within group variability is individual differences among the subjects. One way of reducing error variance is considered in Chapter 7 on factorial designs, and that is by blocking on a variable. One may block on sex, social class, I.Q., etc. All of the variability between blocks is removed from the error term, yielding a more powerful test. In repeated measures designs, blocking is carried to its extreme. We are blocking on each subject. Thus, variability among the subjects due to individual differences is completely removed from the error term. This makes these designs much more powerful than completely randomized designs, where different subjects are randomly assigned to different treatments.

Another distinct advantage of repeated measures designs is that far fewer subjects are required for the study. For example, if three treatments are involved in a completely randomized design, we may require 60 subjects (20 per treatment). With a repeated measures design, we would need only 20 subjects. This can be a very important practical advantage in many cases, since numerous subjects are not readily available in some areas like counseling, school psychology, clinical psychology, and nursing.

Although increased precision and economy of subjects are two distinct advantages of repeated measures designs, these designs have two potentially serious disadvantages, unless care is taken. When several treatments are involved, the *order*

in which treatments are administered might make a difference in the subjects' performance. Thus, it is important to *counterbalance* the order of treatments. For three treatments, counterbalancing involves randomly assigning one third of the subjects to each of the following sequences:

Order of Administration of Treatments		
A	B	C
B	C	A
C	A	B

Another potential disadvantage is the possibility of carryover effects. Thus, it is important to allow sufficient time between treatments to minimize carryover effects, which could occur for example if the treatments were drugs. How much time is necessary is of course a substantive, not a statistical question. Keppel (1983) and Myers (1979) provide further discussion of the two above potential disadvantages.

5.3 SINGLE GROUP REPEATED MEASURES

To illustrate how the variance is partitioned for this simplest design we consider the following data set:

Subjects	Treatments			Means
	1	2	3	
1	30	28	34	30.667
2	14	18	22	18.000
3	24	20	30	24.667
4	38	34	44	38.667
5	26	28	30	28.000
Column Means	26.4	25.6	32	28.000 (grand mean)

We analyze this data in two different ways: (1) as a completely randomized design (pretending there are different subjects for the different treatments), and (2) as a univariate repeated measures analysis. The purpose of including approach 1 is to contrast the error variance that results against the markedly smaller error variance found with the repeated measures design. The reason we mention *univariate* repeated measures analysis is because there is a multivariate approach that can be employed. We discuss and compare the univariate and multivariate approaches after presenting this numerical example.

5.4 COMPLETELY RANDOMIZED DESIGN

This simply involves doing a one way ANOVA, as was done in Chapter 2. Thus, we compute the sums of squares between (SS_b) and the sum of squares within (SS_w):

$$\begin{aligned}
 SS_b &= 5[(26.4 - 28)^2 + (25.6 - 28)^2 + (32 - 28)^2] = 121.6 \\
 SS_w &= (30 - 26.4)^2 + (14 - 26.4)^2 + \cdots + (26 - 26.4)^2 && \text{treatment 1} \\
 &\quad + (28 - 25.6)^2 + (18 - 25.6)^2 + \cdots + (28 - 25.6)^2 && \text{treatment 2} \\
 &\quad + (34 - 32)^2 + (22 - 32)^2 + \cdots + (30 - 32)^2 && \text{treatment 3} \\
 SS_w &= 734.4
 \end{aligned}$$

Now, we need the mean squares:

$$\begin{aligned}
 MS_b &= SS_b / (k - 1) = 121.6 / 2 = 60.80 && \text{and} \\
 MS_w &= SS_w / (N - k) = 734.4 / 12 = 61.20
 \end{aligned}$$

Therefore, $F = MS_b / MS_w = 60.80 / 61.20 = .99$, which is clearly not significant at the .05 level since we have more error variation than effect variation.

5.5 UNIVARIATE REPEATED MEASURES ANALYSIS

Notice that the mean responses for the subjects over the 3 treatments vary considerably (ranging from 18 to 38.667). We quantify this variability through the so-called sum of squares for blocks (SS_{bl}), where here we are blocking on subjects. The error variability that was calculated for the completely randomized analysis is split up into two parts, that is, $SS_w = SS_{bl} + SS_{res}$, where SS_{res} stands for the sum of squares residual. Denote the number of repeated measures by k . Now we calculate the sum of squares for blocks:

$$\begin{aligned}
 SS_{bl} &= k \cdot \sum (\bar{x}_i - \bar{\bar{x}})^2 \\
 &= 3[(30.667 - 28)^2 + (18 - 28)^2 + \cdots + (28 - 28)^2] = 696.02
 \end{aligned}$$

Our error term for the repeated measures analysis is formed by subtracting the sum of squares for blocks from the sum of squares within, $SS_{res} = SS_w - SS_{bl} = 734.4 - 696.02 = 38.38$. Note that the vast majority of the within variability is due to individual differences (696.02 out of 734.4), and that we have removed all of this from our error term. The variability that remains is due to *within subject* variability over treatments. Now,

$$MS_{res} = SS_{res} / (n - 1)(k - 1) = 38.38 / 4(2) = 4.8$$

and our F ratio for the repeated measures analysis is

$$F = MS_b / MS_{res} = 60.80 / 4.8 = 12.67$$

with $(k - 1) = 2$ and $(n - 1)(k - 1) = 4(2) = 8$ degrees of freedom. This is significant at the .05 level (critical value = 4.46), in contrast to the F for the randomized analysis, which was less than 1.

5.6 ASSUMPTIONS IN REPEATED MEASURES ANALYSIS

The three assumptions for a single group univariate repeated measures analysis are:

1. independence of the observations
2. multivariate normality
3. sphericity (sometimes called circularity)

The first two assumptions are also required for the multivariate approach, but the sphericity assumption is not necessary. The reader should recall from Chapter 2 that a violation of the independence assumption is very serious for independent samples ANOVA, and the same holds true for repeated measures analysis. Multivariate normality is somewhat difficult to characterize; however, it does require normality on each of the individual measures. Recall again from Chapter 2 that ANOVA was robust against non-normality. There is also a fair amount of evidence to suggest (Stevens, 1986, p. 207) that MANOVA is also robust against lack of multivariate normality, with respect to type I error.

Before we specify what sphericity means, we wish to note that for many years it was thought that a *stronger* condition called uniformity (compound symmetry) was necessary. The uniformity condition required equality of the population variances for all treatments and also that all population covariances be equal. Schematically for three repeated measures the uniformity condition looks like this:

	1	2	3
1	σ^2	σ_c	σ_c
2	σ_c	σ^2	σ_c
3	σ_c	σ_c	σ^2

In the above, σ^2 represents the common population variance for the three repeated measures, and σ_c represents the common population covariance. Huynh and Feldt (1970) and Rounet and Lepine (1970) independently showed that sphericity is an *exact* condition for the F test to be valid. Sphericity only requires that the variances of the differences for *all pairs* of repeated measures need to be equal.

Sphericity is a weaker condition than uniformity, and defines an additional class of situations where the univariate approach is valid. Consider the covariance matrix below:

$$S = \begin{matrix} & \begin{matrix} y_1 & y_2 & y_3 \end{matrix} \\ \begin{matrix} y_1 \\ y_2 \\ y_3 \end{matrix} & \begin{bmatrix} 1.0 & 0.5 & 1.5 \\ 0.5 & 3.0 & 2.5 \\ 1.5 & 2.5 & 5.0 \end{bmatrix} \end{matrix}$$

The formula for the variance of the difference scores for the i th and j th repeated measures is given by

$$s_{i-j}^2 = s_i^2 + s_j^2 - 2s_{ij}$$

Now we calculate the variance for the differences for each pair of repeated measures

$$\begin{aligned} s_{1-2}^2 &= s_1^2 + s_2^2 - 2s_{12} = 1 + 3 - 2(.5) = 3 \\ s_{1-3}^2 &= 1 + 5 - 2(1.5) = 3 \\ s_{2-3}^2 &= 3 + 5 - 2(2.5) = 3 \end{aligned}$$

The variances are the same for all difference variables, which means the sphericity condition is met, even though uniformity is most definitely not satisfied (all the variances are unequal and the covariances are all unequal).

The multivariate approach to repeated measures is valid for *any* covariance matrix for the repeated measures.

Box (1954) showed that if the sphericity assumption is not met, then the F ratio for the univariate approach is positively biased (we are rejecting falsely too often). In other words, we may set our α level at .05, but may be rejecting falsely 8% or 10% of the time. The extent to which the covariance matrix deviates from sphericity is reflected in a parameter called ϵ (Greenhouse & Geisser, 1959). We give the formula in Exercise 3 for those who are interested. Since $\hat{\epsilon}$ is printed out by SPSS and SAS, there is no need to go through all the tedious calculations. **If sphericity is met, then $\epsilon = 1$, while for the worst possible violation the value of $\epsilon = 1/(k-1)$, where k is the number of repeated measures.** To adjust for the positive bias Greenhouse and Geisser suggest altering the degrees of freedom from

$$(k-1) \text{ and } (k-1)(n-1) \text{ to } 1 \text{ and } (n-1)$$

that is, dividing both degrees of freedom by $(k-1)$.

Doing this makes the test *very* conservative, since adjustment is made for the worst possible case, and we don't recommend it. A more reasonable approach is to estimate ϵ . Then adjust the degrees of freedom from

$(k-1)$ and $(k-1)(n-1)$ to $\hat{\epsilon}(k-1)$ and $\hat{\epsilon}(k-1)(n-1)$

Results from Collier, Baker, Mandeville, and Hayes (1967) and Stoloff (1967) show that this approach keeps the actual α very close to the level of significance.

Huynh and Feldt (1976) found that even multiplying the degrees of freedom by $\hat{\epsilon}$ is somewhat conservative when the true value of ϵ is above about .70. They recommended using the following for those situations:

$$\hat{\epsilon} = \frac{n(i-1)\hat{\epsilon} - 2}{(i-1)[(n-1) - (i-1)\hat{\epsilon}]}$$

The above Huynh epsilon can be printed out by both SPSS MANOVA and SAS GLM.

There are statistical tests for checking sphericity, for example, the Mauchley test presented on SPSS. However, based on the results of Monte Carlo studies (Keselman, Rogan, Mendoza, & Breen, 1980), we don't recommend using these tests.

5.7 SHOULD WE USE THE UNIVARIATE OR MULTIVARIATE APPROACH?

In terms of controlling on type I error, there is no real basis for preferring the multivariate approach, since use of the modified (adjusted) univariate test (i.e., multiplying the degrees of freedom by $\hat{\epsilon}$) yields an honest error rate. The choice then involves a question of power. Now assuming sphericity, the univariate test is more powerful. When sphericity is violated, however, the situation is much more complex. Davidson (1972) has stated, "when small but reliable effects are present with the effects being highly variable ... the multivariate test is far more powerful than the univariate test" (p. 452). And O'Brien and Kaiser (1985), after mentioning several studies that compared the power of the multivariate and modified univariate tests, state, "Even though a limited number of situations has been investigated, this work found that no procedure is uniformly more powerful or even usually the most powerful" (p. 319). Thus, given an exploratory study, we agree with Barcikowski and Robey (1984), who recommend that *both* the univariate and multivariate tests be routinely used because they may differ in the treatment effects that they discern. In such a study half the experimentwise level of significance might be set for each test. Thus, if we wish our overall $\alpha = .05$, simply do each test at $\alpha = .025$.

5.8 COMPUTER ANALYSIS ON SAS AND SPSS
FOR EXAMPLE

In Table 5.1 we present the complete control lines for running the single group repeated measures example given in Section 5.3 on SAS GLM and SPSS MANOVA. Table 5.2 gives the means and standard deviations for the three repeated measures variables, and Table 5.3 presents selected, annotated output from SAS GLM.

TABLE 5.1
SAS and SPSS Control Lines for Single Group Repeated Measures

SAS	SPSS
TITLE 'SINGLE GP REPEATED MEASURES' ;	TITLE 'REPEATED MEASURES' .
DATA SINGLE;	DATA LIST FREE/Y1 Y2 Y3 .
① INPUT SUBJ TREAT REAC @@;	BEGIN DATA.
LINES;	30 28 34 14 18 22 24 20 30
② 1 1 30 1 2 28 1 3 34	38 34 44 26 28 30
2 1 14 2 2 18 2 3 22	END DATA.
3 1 24 3 2 20 3 3 30	LIST.
4 1 38 4 2 34 4 3 44	MANOVA Y1 Y2 Y3/
5 1 26 5 2 28 5 3 30	④ WSFACTOR = TREAT(3) /
PROC PRINT;	WSDSIGN/
PROC GLM;	⑤⑥ ANALYSIS (REPEATED) /
③ CLASS SUBJ TREAT;	PRINT = TRANSFORM CELLINFO (MEANS)
MODEL REAC = SUBJ TREAT;	SIGNIF (UNIV AVERF) / .

① In order to run the single group repeated measures on SAS we treat it as a two way ANOVA, with subjects and treatments as the grouping variables and reaction as the dependent variable.

② The first two numbers of each block of three gives the cell identification. Thus, the first subject in treatment 1 (1 1) had a reaction score of 30, the second subject in treatment 3 (2 3) had a reaction score of 22, etc.

③ In the CLASS statement we list the classification or grouping variables, which here are subject and treatment.

④ The WSFACTOR (within subject factor) and the WSDSIGN (within subject design) are fundamental to running repeated measures analysis on SPSS MANOVA. In the WSFACTOR subcommand we specify which are the repeated measures, or within subject, factors. And we indicate, in parentheses, the number of levels for each repeated measures factor. The WSDSIGN specifies the design on the repeated measures. Here it is simply the treatment effect.

⑤ The TRANSFORM part of the PRINT subcommand prints out in columns the uncorrelated, transformed variables that are created by the program for the multivariate approach (cf. Stevens, 1986, Chapter 13).

⑥ The UNIV is necessary to obtain the significance tests for each of the transformed variables created by the program for the multivariate approach to repeated measures. The AVERF yields the *unadjusted*, overall univariate test for repeated measures.

TABLE 5.2
Means and Standard Deviations for the Drug Data

Placeholder for art from p. 213 of previous edition

5.9 POST HOC PROCEDURES IN REPEATED MEASURES ANALYSIS

As in a one way ANOVA, if an overall difference is found, one would almost always want to determine where the differences lie. This involves a post hoc procedure. There are several reasons for preferring pairwise procedures: (1) they are easily interpreted, (2) they are quite meaningful, and (3) some of these procedures are fairly powerful. The Tukey procedure is appropriate in repeated measures analysis, provided that the sphericity assumption is met. For the drug example this assumption is tenable. We apply the Tukey there, setting overall $\alpha = .05$. Thus, we take at most a 5% chance of one or more false rejections. Recall that we discussed the Tukey procedure in Chapter 2. Remember that the studentized range statistic (denoted by q) is used in the procedure. If there are k groups and the total sample size is N , then any two means are declared significantly different at the .05 level if the following inequality is satisfied:

$$|\bar{x}_i - \bar{x}_j| > q_{.05;k;(n-1)(k-1)} \sqrt{MS_w / n} ,$$

where MS_w is the error term for the one way ANOVA, and n is the common group size. The modification of the Tukey for the one sample repeated measures design is

$$|\bar{x}_i - \bar{x}_j| > q_{.05;k;(n-1)(k-1)} \sqrt{MS_{res} / n} \quad (1)$$

where $(n-1)(k-1)$ is the error degrees of freedom and MS_{res} is the error term, replacing MS_w .

TABLE 5.3
Selected Output From SAS GLM for Single Group Repeated Measures

Placeholder for T0503 from p. 214 of previous edition

TABLE 5.4
Type I Error Rates for the Tukey and Bonferroni Procedures under
Different Violations of the Sphericity Assumption

		$k = 3$		$k = 4$		$k = 5$	
		<i>Tukey</i>	<i>Bonf</i>	<i>Tukey</i>	<i>Bonf</i>	<i>Tukey</i>	<i>Bonf</i>
n	ϵ	min $\epsilon = .50$		min $\epsilon = .33$		min $\epsilon = .25$	
15	1.00	.041	.039	.045	.043	.050	.040
15	.86	.043	.036				
15	.74	.051	.033				
15	.54	.073	.033				
15	.53			.081	.030		
15	.49			.087	.036		
15	.831					.061	.044
15	.752					.067	.042
15	.522					.081	.038
8	.860	.048	.042				
8	.740	.054	.038				
8	.540	.078	.036				
8	.530			.084	.042		
8	.490			.095	.032		
8	.831					.058	.044
8	.752					.060	.042
8	.522					.076	.044

The means, from Table 5.2 are 26.4, 25.6, and 32. If we set overall $\alpha = .05$, then the appropriate studentized range value is $q_{.05; 3, 8} = 4.041$. The error term, from Table 5.3, is 4.8 and the number of subjects is $n = 5$. Therefore, two treatments will be declared significantly different if

$$|\bar{x}_i - \bar{x}_j| > 4.041\sqrt{4.8/5} = 3.96$$

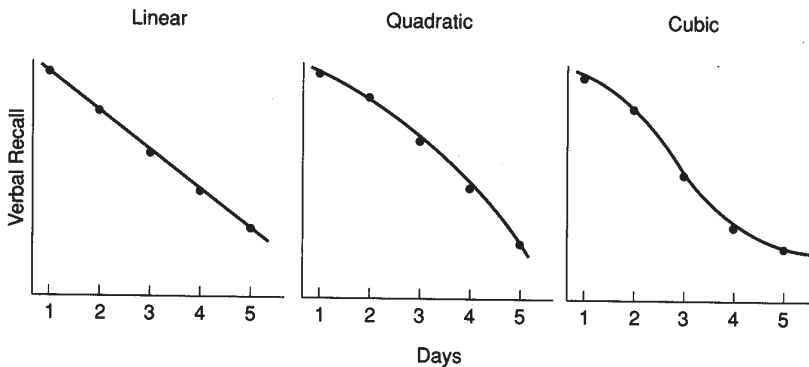
Thus, treatment 3 differs from treatments 1 and 2, but treatments 1 and 2 are not significantly different (as one would have suspected).

There are several other pairwise procedures that Maxwell (1980) discusses in a Monte Carlo study that compared the procedures control on overall α when the sphericity assumption is violated. We present his results for the Tukey and Bonferroni approaches in Table 5.4. The Bonferroni approach in the repeated measures context involves the use of multiple *dependent t* tests. For example, if there are five treatments, then there will be ten paired comparisons. If we wish overall $\alpha = .05$, then we simply do each dependent *t* test at the $.05/10 = .005$ level of significance. Results from Table 5.4 show that the Bonferroni approach keeps the actual $\alpha <$ level of significance in all cases, even when there is a severe violation of the sphericity assumption (e.g., for $k = 3$ the min $\epsilon = .50$ and one of the conditions

modeled had $\epsilon = .54$). Because of this, Maxwell recommended the Bonferroni approach for post hoc pairwise comparisons in repeated measures analysis if the sphericity assumption is violated. Maxwell also studied the power of the five approaches, and found the Tukey to be most powerful. When $\epsilon > .70$ in Table 5.4, the deviation of actual α from the level of significance is less than .02 for the Tukey procedure. This, coupled with the fact that the Tukey tends to be most powerful, would lead us to prefer the Tukey when $\epsilon > .70$. When $\epsilon < .70$, however, then we agree with Maxwell that the Bonferroni approach should be used.

5.10 ONE BETWEEN AND ONE WITHIN FACTOR—A TREND ANALYSIS

We now consider a slightly more complex design, adding a grouping (between) variable. An investigator interested in verbal learning randomly assigns 12 subjects to two treatments. She obtains recall scores on verbal material after 1, 2, 3, 4, and 5 days. Treatments is the grouping variable. She expects there to be a significant effect over time, but wishes a more focused assessment. She wants to mathematically model the form of the decline in verbal recall. For this, trend analysis is appropriate and in particular orthogonal (uncorrelated) polynomials are in order. If the decline in recall is essentially constant over the days, then a significant linear (straight line) trend, or first degree polynomial, will be found. On the other hand, if the decline in recall is slow over the first two days and then drops sharply over the remaining 3 days, a quadratic trend (part of a parabola), or second degree polynomial, will be found. Finally, if the decline is slow at first, then drops off sharply for the next few days and finally levels off, we will find a cubic trend, or third degree polynomial. We illustrate each of these cases below:



The fact that the polynomials are uncorrelated means that the linear, quadratic, cubic, and quartic components are partitioning distinct (different) parts of the variation in the data.

In Table 5.5 we present the SAS and SPSS control lines for running the trend analysis on this verbal recall data. In Chapter 2, in discussing planned comparisons, we indicated that several types of contrasts are available in SPSS MANOVA (Helmert, special, polynomial, etc.), and we also illustrated the use of the Helmert and special contrasts; here the polynomial contrast option is used. Recall these are built into the program, so that all we need do is request them, which is what has been done in the CONTRAST subcommand.

When several groups are involved, as in our verbal recall example, an *additional* assumption is homogeneity of the covariance matrices on the repeated measures for the groups. In our example the group sizes are equal, and in this case a violation of the equal covariance matrices assumption is not serious. That is, the test statistic is robust (with respect to type I error) against a violation of this assumption (cf. Stevens, 1986, Chapter 6). However, if the group sizes are substantially unequal, then a violation is serious, and we indicate in Table 5.5 what should be added to test the assumption.

Table 5.6 gives the means and standard deviations for the two groups on the 5 repeated measures. In Table 5.7 we present selected, annotated output from SPSS MANOVA for the trend analysis. Results from that table show that the groups do not differ significantly ($F = .04, p < .837$) and that there is not a significant group by days interaction ($F = 1.2, p < .323$). There is, however, a quite significant days main effect, and in particular, the LINEAR and CUBIC trends are significant at the .05 level ($F = 239.14, p < .000$ and $F = 10.51, p < .006$, respectively). The linear trend is by far the most pronounced, and a graph of the means for the data in Figure 5.1 shows this, although a cubic curve (with a few bends) will fit the data slightly better.

In concluding this example, the following from Myers (1979) is important:

Trend or orthogonal polynomial analyses should never be routinely applied whenever one or more independent variables are quantitative.... It is dangerous to identify statistical components freely with psychological processes. It is one thing to postulate a cubic component of A , to test for it, and to find it significant, thus substantiating the theory. It is another matter to assign psychological meaning to a significant component that has not been postulated on a priori grounds. (p. 456)

TABLE 5.5
SAS Control Lines and SPSS Command Syntax File for One Between
and One Within Repeated Measures Analysis

SAS	SPSS
TITLE '1 BETW & 1 WITHIN,	TITLE 'ONE BETWEEN AND ONE WITHIN -
DATA TREND;	INTERM. BOOK P. 204'.
INPUT GPID Y1 Y2 Y3 Y4 Y5;	DATA LIST FREE/GPID Y1 Y2 Y3 Y4 Y5.
CARDS;	BEGIN DATA.
1 26 20 18 11 10	1 26 20 18 11 10 1 34 35 29 22 23
1 34 35 29 22 23	1 41 37 25 18 15
1 41 37 25 18 15	1 29 28 22 15 13 1 35 34 27 21 17
1 29 28 22 15 13	1 28 22 17 14 10
1 35 34 27 21 17	1 38 34 28 25 22 1 43 37 30 27 25
1 28 22 17 14 10	2 42 38 26 20 15 2 31 27 21 18 13
1 38 34 28 25 22	2 45 40 33 25 18
1 43 37 30 27 25	2 29 25 17 13 8 2 29 32 28 22 18
2 42 38 26 20 15	2 33 30 24 18 7
2 31 27 21 18 13	2 34 30 25 24 23 2 37 31 25 22 20
2 45 40 33 25 18	END DATA.
2 29 25 17 13 8	LIST.
2 29 32 28 22 18	MANOVA Y1 TO Y5 BY GPID (1,2)/
2 33 30 24 18 7	③ WSFACTOR = DAY(5)/
2 34 30 25 24 23	④ CONTRAST(DAY) = POLYNOMIAL/
2 37 31 25 22 20	③ WSDESIGN = DAY/
PROC GLM;	⑤ RENAME = MEAN, LINEAR, QUAD, CUBIC, QUART/
CLASS GPID;	PRINT = TRANSFORM CELLINFO(MEANS)
MODEL Y1 Y2 Y3 Y4 Y5 = GPID;	SIGNIF(AVERF)/
① REPEATED DAY 5 (12 3 4 5)	ANALYSIS(REPEATED)/
② POLYNOMIAL/SUMMARY;	⑥ DESIGN = GPID/.

① The REPEATED statement is fundamental for running repeated measures designs on SAS. The general form is REPEATED factorname levels (level values) transformation/options; Note that the level values are in parentheses. We are interested in polynomial contrasts on the repeated measures, and so that is what has been requested. Other transformations are available HELMERT, PROFILE, etc.—see SAS USER'S GUIDE: STATISTICS, Version 5, p. 454).

② SUMMARY here produces ANOVA tables for each contrast defined by the within subject actors.

③ Recall again that the WSFACTOR (within subject factor) and the WSDESIGN (within subject design) subcommands are fundamental for running multivariate repeated measures analysis on SPSS.

④ If we wish trend analysis on the DAY repeated measure variable, then all we need do is request POLYNOMIAL on the CONTRAST subcommand.

⑤ In this RENAME subcommand we are giving meaningful names to the polynomial contrasts being generated.

⑥ It is important to realize that with SPSS MANOVA there is a design subcommand WSDESIGN) for the within or repeated measures factor(s) and a *separate* DESIGN subcommand or the between(grouping) factor(s).

TABLE 5.6
Means and Standard Deviations for One Between and One Within
Repeated Measures

Placeholder for T0506 on p. 219 of previous edition

5.11 POST HOC PROCEDURES FOR THE ONE BETWEEN AND ONE WITHIN DESIGN

In the one between and one within, or mixed model, repeated measures design, we have both the assumption of sphericity *and* homogeneity of the covariance matrices for the different levels of the between factor. This combination of assumptions has been called multisample sphericity. Keselman and Keselman (1988) conducted a Monte Carlo study examining how well four post hoc procedures controlled overall alpha under various violations of multisample sphericity. The four

TABLE 5.7
Selected Output from SPSS for One Between and One Within

Insert T0507 from p. 220 of previous edition

① The group and group by days interaction are not significant, although the unadjusted DAYS main effect is significant at the .05 level.

② The last four columns of numbers are the coefficients for orthogonal polynomials, although they may look strange since each column is scaled such that the sum of the squared coefficients equals 1. Textbooks typically present the coefficients for 5 levels as follows:

Linear	-2	-1	0	1	2
Quadratic	2	-1	-2	-1	2
Cubic	-1	2	0	-2	1
Quartic	1	-4	6	-4	1

Compare, for example, *Fundamentals of Experimental Design*, Myers, 1979, p. 548.

③ This value of $\hat{\epsilon}$ indicates a severe violation of the sphericity assumption, although the adjusted univariate test is still easily significant at the .05 level.

(Continued)

TABLE 5.7
(Continued)

Placeholder for T0507 from p. 221 of previous edition

TABLE 5.7
(Continued)

Insert from p. 222 of previous edition.

procedures were: the Tukey, a modified Tukey employing a nonpooled estimate of error, a Bonferroni t statistic, and a t statistic with a multivariate critical value. These procedures were also used in the Maxwell (1980) study of post hoc procedures for the single group repeated measures design.

Keselman and Keselman set the number of groups at 3 and considered 4 and 8 levels for the within (repeated) factor. They considered both equal and unequal group sizes for the between factor. Recall that ϵ quantifies departure from sphericity, and $\epsilon = 1$ means sphericity, with $1/(k-1)$ indicating maximum departure from sphericity. They investigated $\epsilon = .75$ (a relatively mild departure) and $\epsilon = .40$ (a severe departure for the 4 level case, given the minimum value there would be .33). Selected results from their study are presented below for the four level within factor case.

		Tukey(pooled)	Bonferroni	Multivariate
$\epsilon = .75$	equal covariance matrices & gp sizes	6.34	3.46	1.70
	unequal covariance matrices, but equal group sizes	7.22	4.32	2.48
	unequal covariance matrices and gp sizes—larger variability with smaller group size	14.78	11.38	7.04
$\epsilon = .40$	equal covariance matrices & gp sizes	11.36	2.38	1.16
	unequal covariance matrices, but equal group sizes	10.08	2.70	1.56
	unequal covariance matrices and gp sizes—larger variability with smaller group size	17.80	6.34	3.94

The group sizes for the values presented above were 13, 10, and 7. The entries in the body of the table are to be compared against an overall alpha of .05.

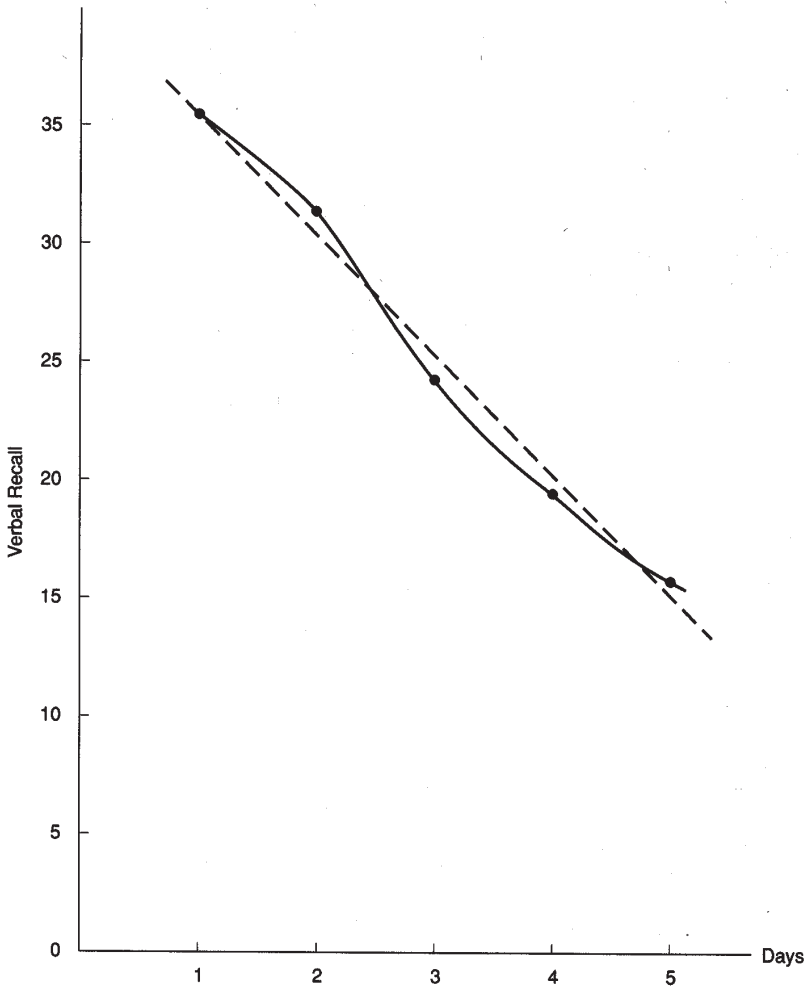


FIGURE 5.1 Linear and Cubic Plots for Verbal Recall Data

The above results show that the Bonferroni approach keeps the overall alpha less than .05, provided you do not have *both* unequal group sizes and unequal covariance matrices. If you want to be confident that you will be rejecting falsely no more than your level of significance, then this is the procedure of choice. In my opinion, the Tukey procedure is acceptable for $\epsilon = .75$, as long as there are *equal*

group sizes. For the other cases, the error rates for the Tukey are at least double the level of significance, and therefore not acceptable.

Recall that the pooled Tukey procedure for the single group repeated measures design was to reject if

$$|\bar{x}_i - \bar{x}_j| > q_{.05;k;(n-1)(k-1)} \sqrt{MS_{res} / n}$$

where n is the number of subjects, k is the number of levels, and MS_{res} is the error term (Equation 1).

For the one between and one within design with J groups and k within levels, we declare two marginal means (means for the repeated measures levels over the J groups) different if

$$|\bar{x}_i - \bar{x}_j| > q_{.05;k;(N-J)(k-1)} \sqrt{MS_{kcs} / J / N}$$

where the mean square is the within subjects error term for the mixed model and N is total number of subjects.

5.12 ONE BETWEEN AND TWO WITHIN FACTORS

We now consider the repeated measures analysis for a one between and two within design, using data from Elashoff (1981). Two groups of subjects are given three different doses of two drugs. There are several different questions of interest in this study. Will the drugs be differentially effective for different groups? Is the effectiveness of the drugs dependent on dose level? Is the effectiveness of the drugs dependent on both dose level and on the group?

The design for this study is given below schematically:

		Drug 1			Drug 2		
Dose		D_1	D_2	D_3	D_1	D_2	D_3
Group 1	S_1	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
	S_2						
	S_8						
Group 2	S_9						
	S_{10}						
	S_{16}						

Note that there are six measures for each subject (Y_1 to Y_6). Also, we have a crossed design on the within variables of drug and dose.

The complete control lines for running this analysis on SAS are given in Table 5.8. The means and standard deviations for the six repeated measures are given in Table 5.9. Table 5.10 presents the univariate analyses for this design from the SAS GLM program. Although the control lines for SAS in Table 5.8 yield *both* the

TABLE 5.8
SAS Control Lines for One Between and Two Within Repeated Measures

```

SAS
TITLE 'ELASHOFF DATA';
DATA ELAS;
INPUT GP Y1 Y2 Y3 Y4 Y5 Y6;
CARDS;
1 19 22 28 16 26 22
1 11 19 30 12 28 28
1 20 24 24 24 22 29
1 21 25 25 15 10 26
1 18 24 29 19 26 28
1 17 23 28 15 23 22
1 20 23 23 26 21 28
1 14 20 29 25 29 29
2 16 20 24 30 34 36
2 26 26 26 24 30 32
2 22 27 23 33 36 45
2 16 18 29 27 26 34
2 19 21 20 22 22 21
2 20 25 25 29 29 33
2 21 22 23 27 26 35
2 17 20 22 23 26 28
PROC GLM;
CLASS GP;
① MODEL Y1 Y2 Y3 Y4 Y5 Y6 = GP;
② REPEATED DRUG 2, DOSE 3;

```

① Recall that in the MODEL statement the dependent variables, the repeated measures here, go on the left side and the classification or grouping variable(s) go on the right side.

② When there is more than one repeated measures factor, they *must* be separated by a comma, and the product of the levels of all factors must equal the number of dependent variables in the MODEL statement.

univariate and multivariate analyses, we have just presented the univariate analyses because the Greenhouse–Geisser epsilons in Table 5.10 are greater than .70. For such a relatively mild violation of sphericity, it has been shown that the type I error rate remains at essentially the level of significance.

Results from Table 5.10 show that we have significant drug, group, and dose main effects and a significant drug by group interaction at the .05 level. To ascertain what was responsible for the drug, group and drug by group interactions we take the means from Table 5.9 and insert them into the design, yielding:

TABLE 5.9
Means and Standard Deviations for One Between
and Two Within Repeated Measures

Placeholder for T0509 from p. 227 of previous edition

TABLE 5.10
Univariate Analyses from SAS GLM for One Between and Two Within

Placeholder for T0510 parts a & b from pp. 228-229 of previous edition
--

Placeholder for art p. 229 of previous edition.

-
- ① Since both $\hat{\epsilon}$ s are $> .70$, the univariate approach is preferred, since the type I error rate is controlled and it may be more powerful than the multivariate approach.
- ② Groups differ significantly at the $.05$ level, since $.0185 < .05$.
- ③ & ④ The drug main effect and drug by group interaction are significant at the $.05$ level, while the dose main effect is also significant at the $.05$ level.
- ⑤ Note that 4 different error terms are involved in this design, an additional complication with complex repeated measures designs. The error terms are boxed.

Dose	Drug					
	1	2	3	1	2	3
Group 1	17.5	22.5	27	19	21.88	26.5
Group 2	19.63	22.38	24	26.88	28.63	33.00

Now, collapsing on dose, the group by drug means are obtained:

	Drug	
	1	2
Group 1	22.3	22.46
Group 2	22.00	29.50

The mean in cell 11 (22.33) is simply the average of 17.5, 22.5, and 27, the mean in cell 12 (22.46) is the average of 19, 21.88, and 26.5, etc. Now it is apparent that the “outlier” cell mean of 29.50 is what was responsible for both main effects and the interaction being significant. Note that if this cell mean were about 22 or 23, as the others, then none of the effects would have been significant.

Now, we obtain the level means for DOSE and then apply the Tukey procedure to see which dose levels differ significantly. The dose level means are 20.753, 23.848, and 27.625. Now, two DOSE level means will differ significantly if

$$|\bar{x}_i - \bar{x}_j| > q_{.05;3,28} \sqrt{10.391/16} ,$$

where 10.391 is the mean square error term for DOSE (cf. Table 5.9), 16 is the number of subjects for each dose, and 28 is the error degrees of freedom. Calculation yields:

$$|\bar{x}_i - \bar{x}_j| > 3.486 \sqrt{10.391/16} = 2.809$$

Since the smallest difference between any two level means is 3.095 for levels 1 and 2, this means that all dose levels differ significantly from one another.

One Between and Two Within on SPSS for Windows 12.0

Once the data are in the editor, click on ANALYZE and scroll down to GENERAL LINEAR MODEL. At this point the screen appears as at the top of Table 5.11. When you scroll across to GLM-REPEATED MEASURES and click the screen at the left middle of Table 5.11 appears. Click within the WITHIN SUBJECT FACTOR NAME box and type in drug. Then click within the NUMBER OF LEVELS box and type in 2. The ADD box will light up; click on it. Do the same for DOSE (remember DOSE has 3 levels), click on ADD, and at this point the screen will appear as at the right middle in Table 5.11. When you click on DEFINE, the screen at the bottom of Table 5.11 appears. Click on y 1 and then click on the forward arrow to put y1 in position 1,1. Do the same for y2

through y6. Finally, click on GP and click on the forward arrow to make GP a between subjects factor. The OK box will light up. Simply click on OK to run the analysis.

5.13 TOTALLY WITHIN DESIGNS

There are research situations where the *same* subjects are measured under various treatment combinations, that is, where the same subjects are in each cell of the design. This may particularly be the case when not many subjects are available. We consider three examples to illustrate:

Example 1

A researcher in child development is interested in observing the same group of pre-school children (all 4 years of age) in two situations at two different times (morning and afternoon) of the day. She is concerned with the extent of their social interaction, and will measure this by having two observers independently rate the amount of social interaction. The average of the two ratings will serve as the dependent variable. The within factors here are situation and time of day.

There are 4 scores for each child: social interaction in situation 1 in the morning and afternoon, and social interaction in situation in the morning and afternoon. We denote the four scores by Y1, Y2, Y3, and Y4.

Such a totally within repeated measures design is easily setup on SPSS MANOVA. The command syntax file is given below:

```
TITLE 'TWO WITHIN DESIGN' .
DATA LIST FREE/Y1 Y2 Y3 Y4 .
BEGIN DATA.
DATA LINES
END DATA.
MANOVA Y1 TO Y4 /
WSFACTOR = SIT(2),TIME(2) /
WSDSIGN /
PRINT = TRANSFORM CELLINFO (MEANS) /
ANALYSIS (REPEATED) / .
```

Example 2

A social psychologist is interested in determining how self-reported anxiety level for 35–45 year old men varies as a function of situation, who they are with, and how many people are involved. A questionnaire will be administered to 20 such men, asking them to rate their anxiety level (on a Likert scale from 1 to 7) in 3 situ-

TABLE 5.11
SPSS for Windows 12.0 Screens for for One Between and Two Within
Repeated Measures

Repeated Measures Define Factor(s)

Within-Subject Factor Name: factor1 Define

Number of Levels: 1 Reset

Add Change Remove Cancel Help Measure >>

1

Repeated Measures Define Factor(s)

Within-Subject Factor Name: drug Define

Number of Levels: 2 Reset

Add Change Remove Cancel Help Measure >>

2

Repeated Measures Define Factor(s)

Within-Subject Factor Name: dose Define

Number of Levels: 3 Reset

Add drug(2) Change Remove Cancel Help Measure >>

3

Repeated Measures Define Factor(s)

Within-Subject Factor Name: Define

Number of Levels: Reset

Add drug(2) dose(3) Change Remove Cancel Help Measure >>

4

Repeated Measures

Within-Subjects Variables: (drug.dose) OK

d1dose1 d1dose2 d1dose3 d2dose1 d2dose2 d2dose3

Between-Subjects Factor(s): Paste Reset Cancel Help

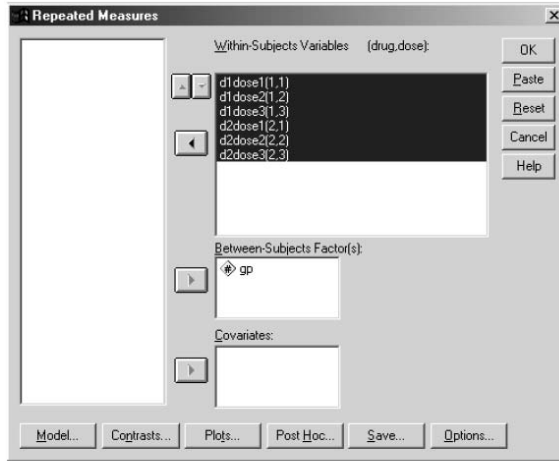
Covariates:

Model... Contrasts... Plots... Post Hoc... Save... Options...

5

Use ANALYZE—GENERAL LINEAR MODEL—REPEATED MEASURES to get to screen 1. Click on ADD to go from screen 3 to screen 4. Click on DEFINE to go from screen 4 to screen 5. Click on OK in screen 6 to run the analysis.

TABLE 5.11
SPSS for Windows 12.0 Screens for for One Between and Two Within
Repeated Measures (*continued*)



6

ations (going to the theater, going to a football game, and going to a dinner party), with primarily friends and primarily strangers, and with a total of 6 people and with 12 people. Thus, the men will be reporting anxiety for 12 different contexts. This is a three within, crossed repeated measures design, where situation (3 levels) is crossed with nature of group (2 levels) and with number in group (2 levels).

Example 3

Suppose in an ergonomic study we are interested in the effects of day of the work week and time of the day (AM or PM) on various measures of posture. We select 30 computer operators and for this example we consider just one measure of posture called shoulder flexion. We then have a two factor totally within design which looks as follows:

	Monday		Wednesday		Friday	
	AM	PM	AM	PM	AM	PM
1						
2						
3						
.						
.						
.						
30						

TABLE 5.12
SPSS 12.0 Command Syntax File for Helmert Contrasts
for a Repeated Measures Factor

TITLE 'HELMERT CONTRASTS FOR REPEATED MEASURES'.												
DATA LIST FREE/Y1 Y2 Y3 Y4.												
BEGIN DATA.												
6.6	1.3	2.5	2.1	3.0	1.4	3.8	4.4	4.7	4.5	5.8	4.7	
6.2	6.1	6.1	6.7	3.2	6.6	7.6	8.3	2.5	6.2	8.0	8.2	
2.8	3.6	4.4	4.3	1.1	1.1	5.7	5.8	2.9	4.9	6.3	6.4	
5.5	4.3	5.6	4.8									
END DATA.												
LIST.												
① MANOVA Y1 TO Y4/												
WSFACTOR = DRUGS(4) /												
② CONTRAST (DRUGS) = HELMERT/												
WSDESIGN = DRUGS/												
③ RENAME = MEAN, HELMERT1, HELMERT2, HELMERT3/												
PRINT = CELLINFO (MEANS) TRANSFORM/												
ANALYSIS (REPEATED) /.												

- ① Recall that the four repeated measures are treated as 4 dependent variables.
- ② Since HELMERT is one of the standard set of contrasts available in SPSS MANOVA, all we need do is request it in the CONTRAST subcommand.
- ③ We have simply given meaningful names to the transformed variables. The first transformed variable is a general mean, and the last 3 transformed variables are the Helmert contrasts that we wish to test the significance of.

5.14 PLANNED COMPARISONS IN REPEATED
MEASURES DESIGNS

Planned comparisons can be easily set up on SPSS MANOVA or on SAS GLM for repeated measures factors. To illustrate, we consider data from a study reported by Bock (1975). The study involved the effect of three drugs on the duration of sleep for 10 mental patients. The drugs were given orally on alternate evenings, and the hours of sleep were compared with an intervening control night. Each of the drugs was tested a number of times with each patient. The average number of hours of sleep was the dependent measure. Schematically, we have

	Control	Drug Type I	Drug Type II	Drug Type III
Subjects 1				
2				
.				
.				
.				
10				

Drug Type I was distinctly different from the remaining two drugs, which were somewhat similar in composition. Three relevant questions here are:

1. Does drug have a different effect on duration of sleep than no drug?
2. Does drug Type I produce a different effect from types II and III?
3. Do drug types II and III, which are similar, have a differential effect on sleep?

These questions correspond to the following contrasts on the repeated measures:

	y_1	y_2	y_3	y_4
L_1	1	-.33	-.33	-.33
L_2	0	1	-.50	-.50
L_3	0	0	1	-1

Notice in the above that each level of the repeated measure is contrasted against the *average* of the remaining levels. This kind of set of contrasts are called Helmert contrasts. They are built into the SPSS and SAS packages. All one need do is request them. In Table 5.12 we present the complete command syntax for running the Helmert contrasts on SPSS MANOVA.

5.15 SUMMARY

1. Repeated measures designs are more powerful than completely randomized designs, since the variability due to individual differences is removed from the error term, and individual differences are the major reason for error variance.
2. Two major advantages of repeated measures designs are increased precision (because of the smaller error term), and economy of subjects. Two potential disadvantages are that the order of treatments may make a difference (this can be dealt with by counterbalancing) and carryover effects.
3. Either a univariate or multivariate approach can be used for repeated measures analysis. If the sphericity assumption is tenable, then the univariate approach is preferred as it is more powerful.
4. If sphericity is violated, then the type I error rate for the univariate approach is inflated. However, a modified univariate approach (obtained by multiplying each of the degrees of freedom by $\hat{\epsilon}$) yields an honest type I error rate.
5. *As both the modified univariate and multivariate approaches control type I error, the choice between them involves the issue of power.* To keep things simpler in this text, I simply illustrate and use the modified univariate ap-

proach. However, as I point out in my multivariate text, neither approach is even usually more powerful, and therefore I recommend there that both approaches should be used, since they may differ in the effects they will discern.

6. If sphericity is tenable, then the Tukey is a good post hoc procedure for locating pairwise differences. If sphericity is not tenable, then the Bonferroni approach should be used. That is, do multiple correlated t tests, but use the Bonferroni Inequality to keep the overall α level under control.
7. When several groups are involved, then an additional assumption is homogeneity of the covariance matrices for the groups. This can be checked with the Box test, and would be of most concern when the group sizes are sharply unequal.

EXERCISES

1. Consider the following data for a single group repeated measures with 8 subjects measured for 4 treatments:

Subjects	Treatments			
	1	2	3	4
1	5	6	2	5
2	3	4	1	6
3	3	7	4	10
4	6	8	3	3
5	4	9	7	8
6	5	7	4	9
7	2	10	1	2
8	4	3	2	5

- (a) Do a univariate repeated measures analysis on this data, testing for significance at the .05 level.
 - (b) Use the Tukey post hoc procedure to locate the significant pairwise differences at the .05 level.
 - (c) Run the above data on SPSS MANOVA to check your results.
2. Give an example or two where a two or three within subjects design would make sense. As a starter for you, consider driving behavior (measured by number of steering errors) or smoking behavior (number of cigarettes smoked) for a group of subjects, and a few key factors that you think might influence such behavior.

3. Output from SPSS MANOVA for the single sample (5 subjects and 3 levels) repeated measures design in Table 5.1 includes the following:

GREENHOUSE-GEISSER EPSILON = .66564

HUYNH-FELDT EPSILON = .87240

LOWER-BOUND EPSILON = .50000

The covariance matrix for the three measures is

$$\mathbf{S} = \begin{bmatrix} 76.8 & 53.2 & 69.0 \\ 53.2 & 42.8 & 47.0 \\ 69.0 & 47.0 & 64.0 \end{bmatrix}$$

The formula for the Greenhouse–Geisser epsilon is:

$$\hat{\epsilon} = \frac{k^2(\bar{s}_{ii} - \bar{s})^2}{(k-1)(\sum \sum s_{ij}^2 - 2k \sum_i \bar{s}_i^2 + k^2 \bar{s}^2)}$$

where

\bar{s} is the mean of all entries in the covariance matrix \mathbf{S}

\bar{s}_{ii} is mean of entries on main diagonal of \mathbf{S}

\bar{s}_j is mean of all entries in row i of \mathbf{S}

s_{ij} is ij th entry of \mathbf{S}

- Using this formula, verify the SPSS value of .66564.
- Using the equation given in the chapter relating the Greenhouse–Geisser and the Huynh–Feldt epsilons, verify the value of .87240.
- Why is the LOWER-BOUND EPSILON value given as .500?
- For the one between and one within design in Table 5.7, the Greenhouse–Geisser epsilon = .44629 and Huynh–Feldt epsilon = .54366 on the SPSS printout. A generalized formula relating these measures for this design is given by

$$\hat{\epsilon} = \frac{ng(k-1)\hat{\epsilon} - 2}{(k-1)[g(n-1) - (k-1)\hat{\epsilon}]}$$

where g is the number of groups, n is the number of subjects per group, and k is the number of levels for the within variable. Using this relationship, show how the Huynh–Feldt value of .54366 follows from the Greenhouse–Geisser value of .44629.

4. Consider the following hypothetical data from a study comparing the relative efficacy of a behavior modification approach to dieting vs. a behavior

modification approach + exercise on weight loss for a group of overweight women. There is also a control group. First, 18 women who are between 20 and 30 years old are randomly assigned to one of the three groups. Then, six each of women 30 to 40 years old are randomly assigned to one of the three groups. The investigator wishes to determine whether age might moderate the effect of the diet approaches. The weight loss for the women is measured two months, four months, and six months after the diets begin. Thus, we have a two between and one within repeated measures design.

	DIET	AGE	WGTLOSS1	WGTLOSS2	WGTLOSS3
CONTROL 20–30 YRS	1.00	1.00	4.00	3.00	3.00
	1.00	1.00	4.00	4.00	3.00
	1.00	1.00	4.00	3.00	1.00
	1.00	1.00	3.00	2.00	1.00
	1.00	1.00	5.00	3.00	2.00
	1.00	1.00	6.00	5.00	4.00
CONTROL 30–40 YRS	1.00	2.00	6.00	5.00	4.00
	1.00	2.00	5.00	4.00	1.00
	1.00	2.00	3.00	3.00	2.00
	1.00	2.00	5.00	4.00	1.00
	1.00	2.00	4.00	2.00	2.00
	1.00	2.00	5.00	2.00	1.00
BEH. MOD 20–30 YRS	2.00	1.00	6.00	3.00	2.00
	2.00	1.00	5.00	4.00	1.00
	2.00	1.00	7.00	6.00	3.00
	2.00	1.00	6.00	4.00	2.00
	2.00	1.00	3.00	2.00	1.00
	2.00	1.00	5.00	5.00	4.00
BEH. MOD 30–40 YRS	2.00	2.00	4.00	3.00	1.00
	2.00	2.00	4.00	2.00	1.00
	2.00	2.00	6.00	5.00	3.00
	2.00	2.00	7.00	6.00	4.00
	2.00	2.00	4.00	3.00	2.00
	2.00	2.00	7.00	4.00	3.00
BEH. MOD. + EXER. 20–30 YRS	3.00	1.00	8.00	4.00	2.00
	3.00	1.00	3.00	6.00	3.00
	3.00	1.00	7.00	7.00	4.00
	3.00	1.00	4.00	7.00	1.00
	3.00	1.00	9.00	7.00	3.00
	3.00	1.00	2.00	4.00	1.00
BEH. MOD. + EXER. 30–40 YRS	3.00	2.00	3.00	5.00	1.00
	3.00	2.00	6.00	5.00	2.00
	3.00	2.00	6.00	6.00	3.00
	3.00	2.00	9.00	5.00	2.00
	3.00	2.00	7.00	9.00	4.00
	3.00	2.00	8.00	6.00	1.00

- (a) Run the analysis on SPSS MANOVA, obtaining both the multivariate and univariate results.
 - (b) Which of the between effects are significant at the .05 level?
 - (c) Given the values of the Greenhouse–Geisser and Huynh–Feldt epsilons, would the univariate or multivariate approach be preferred?
 - (d) Which of the within effects are significant at the .05 level?
 - (e) Using the appropriate means (cell, row or column), interpret the results.
5. Run the Helmert planned comparisons given in Table 5.12 on SPSS MANOVA. If overall alpha is set at .10, then which are significant? What do the significant contrasts represent?
6. Recall that in the Elashoff data example in section 5.12 two groups of subjects were given three different doses of two drugs, which yielded a one between and two within repeated measures design. Suppose that the two groups of subjects had been given the different doses of the drugs under two different conditions. Then we would have a one between and three within design. Show the SPSS MANOVA control lines for running this analysis.
7. Consider the following data. The dependent variable is Beck depression score:

	WINTER	SPRING	SUMMER	FALL
1	7.50	11.55	1.00	1.21
2	7.00	9.00	5.00	15.00
3	1.00	1.00	.00	.00
4	.00	.00	.00	.00
5	1.06	.00	1.10	4.00
6	1.00	2.50	.00	2.00
7	2.50	.00	.00	2.00
8	4.50	1.06	2.00	2.00
9	5.00	2.00	3.00	5.00
10	2.00	3.00	4.21	3.00
11	7.00	7.35	5.88	9.00
12	2.50	2.00	.01	2.00
13	11.00	16.00	13.00	13.00
14	8.00	10.50	1.00	11.00

- (a) Run this on SPSS or SAS as a single group repeated measures. Is it significant at the .05 level, assuming sphericity?

8. A researcher is interested in the smoking behavior of a group of 30 men, 10 of which are 30–40, 10 are 41–50, and the remaining 10 are 51–60. She wishes to determine if how much they smoke is influenced by the time of the day (morning or afternoon) and by context (at home or in the office). The men are observed in each of the 4 situations and the number of cigarettes is recorded. She also wishes to determine whether the age of the men influences their smoking behavior.
- What type of a repeated measures design is this?
 - Show the complete SPSS MANOVA control lines (put DATA for the data lines) for running the analysis.
9. Consider the following data set:

TREATMENTS		
1	2	3
5	6	1
3	4	2
3	7	1
6	8	3
6	9	3
4	7	2
5	9	2

- Do a single group repeated measures analysis on this data.

Simple and Multiple Regression

CONTENTS

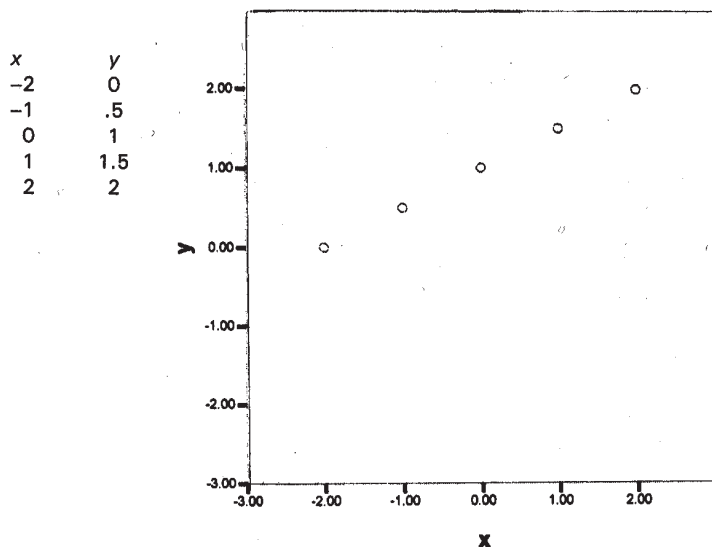
- 6.1 Simple Regression
- 6.2 Assumptions for the Errors
- 6.3 Influential Data Points
- 6.4 Multiple Regression
- 6.5 Breakdown of Sum of Squares in Regression and F Test for Multiple Correlation
- 6.6 Relationship of Simple Correlations to Multiple Correlation
- 6.7 Multicollinearity
- 6.8 Model Selection
- 6.9 Two Computer Examples
- 6.10 Checking Assumptions for the Regression Model
- 6.11 Model Validation
- 6.12 Importance of the Order of Predictors in Regression Analysis
- 6.13 Other Important Issues
- 6.14 Outliers and Influential Data Points
- 6.15 Further Discussion of the Two Computer Examples
- 6.16 Sample Size Determination for a Reliable Prediction Equation
- 6.17 ANOVA as a Special Case of Regression Analysis
- 6.18 Summary of Important Points
- Appendix The PRESS Statistic

6.1 SIMPLE REGRESSION

Here we are predicting a dependent (outcome) variable from a single predictor. Several examples come to mind. One may wish to predict chemistry achievement

from I.Q. One may wish to predict a person's heart rate from blood pressure. A farmer may wish to predict yield from level of the fertilizer.

Before we get into simple regression, let us review some basic concepts from high school. Recall that in high school you may have been told to graph an equation such as $y = 1 + .5x$. To do so you take a range of x values, determine the corresponding y values, and then plot the points. It would look like this:

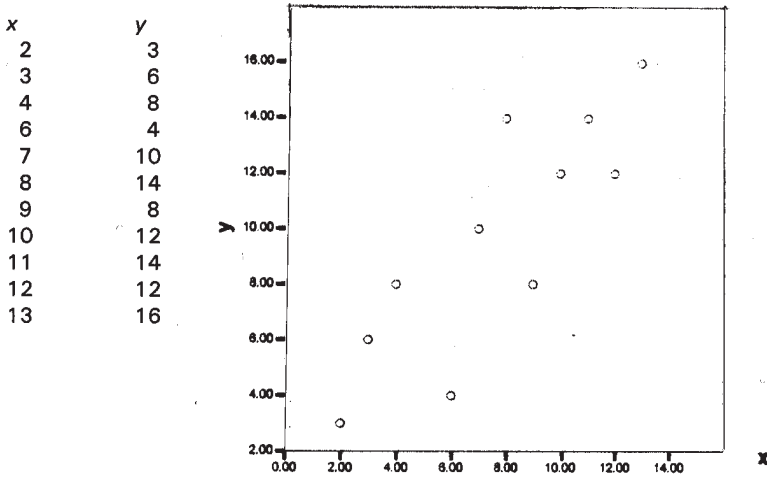


When you did this each y value was *exactly* determined once the x value was specified. You probably did not think a lot about that fact. This is called a deterministic model. When we collect data and are attempting to predict some y (like college GPA) from a single x (like high school GPA), it should be obvious that perfect prediction is not going to happen. Why? Because there are many other factors that determine college GPA, like what your major is, where you go to college, boyfriends and girlfriends, a death in the family, a divorce, attitude toward school, etc. Because of all these other factors we need to set up what is called a *probabilistic* model, where we allow for error in prediction. We will assume a linear relationship between y and x . The probabilistic model is as follows:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, 2, \dots, n$$

The part I have underlined corresponds to the linear relationship, and the other part is the error of prediction.

To illustrate the above, consider the following data set and scatterplot:



From the above plot it is obvious that a straight line will not fit the points perfectly. Yet it is also clear that as x increases y increases and that the relationship seems primarily linear. We wish to model this linear relationship, and we will see that a “least squares” regression line does a pretty good job.

We consider two examples to illustrate simple regression. The first example uses artificial data for a small data set. The second example, based on real data, provides a more realistic use of simple regression in practice.

Before we get into the examples, let us consider more precisely what is done in simple regression. First, we are assuming a *linear* relationship exists between the dependent variable and the predictor. This means there is a significant correlation between x and y . We are modeling y , assuming it is linearly related to x (predictor). The mathematical model looks like this:

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad i = 1, 2, \dots, n$$

where β_0 and β_1 are to be estimated, and the e_i are the errors of prediction. There are assumptions concerning the e_i which we get to later. How do we estimate the β s? The *least squares* criterion is used; that is, the sum of the squared estimated errors of prediction is minimized.

$$\hat{e}_1^2 + \hat{e}_2^2 + \dots + \hat{e}_n^2 = \sum_{i=1}^n \hat{e}_i^2 = \min$$

Now, $\hat{e}_i = y_i - \hat{y}_i$ where y_i is the actual score on the dependent variable and \hat{y}_i is the estimated score for the i th subject.

The score for each subject defines a point in the plane. What the least squares criterion does is find the line that best fits the points. Geometrically this corre-

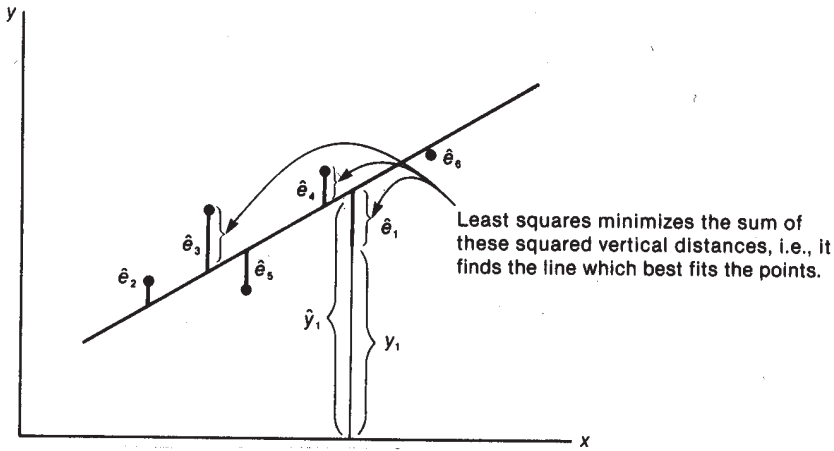


FIGURE 6.1 Geometrical Representation of Least Squares Criterion

sponds to *minimizing* the sum of the squared vertical distances (\hat{e}_i^2) of each subject's score from their estimated score. This is illustrated in Figure 6.1.

Example 1

The above is abstract. To give the reader a feel for the errors of prediction and just plotting of points, we consider our first example with just 11 data points. In Table 6.1 we present selected printout from the SPSS for Windows 12.0 regression run of this data. First, the correlation of .841 shows that there is a strong linear relationship. Second, from the unstandardized coefficients we can construct the prediction equation. We have put that equation on the diagram below. Third, in Table 6.1 we have the unstandardized predicted values and errors. In Figure 6.2 we present the regression line, along with geometric illustrations of what some of the estimated y values look like and how the errors of prediction are obtained by simply taking the difference between the person's actual y score and the person's predicted score.

Example 2

For our second example, using real data, we consider part of a Sesame Street database from Glasnapp and Poggio (1985), who present data on many variables, including 12 background variables and 8 achievement variables, for 240 subjects. Sesame Street was developed as a television series aimed mainly at teaching preschool skills to three to five year old children. Data was collected on many achievement variables both before (pretest) and after (posttest) viewing of the series. We

TABLE 6.1
Selected Simple Regression Output From SPSS for Windows 12.0

Placeholder for T0601 on p. 243 of previous edition.

TABLE 6.2
SPSS Command Syntax for Simple Regression on Sesame Street Data
and Selected Printout

TITLE 'SIMPLE REGRESSION ON SESAME DATA'.	
DATA LIST FREE/PREBODY POSTBODY.	
BEGIN DATA.	
DATA LINES	
END DATA.	
REGRESSION DESCRIPTIVES = DEFAULT/	
VARIABLES = PREBODY POSTBODY/	
DEPENDENT = POSTBODY/	
①	METHOD = ENTER/
②	SCATTERPLOT (POSTBODY, PREBODY) /.

① DESCRIPTIVES = DEFAULT subcommand yields the means, standard deviations and the correlation matrix for the variables.

② This SCATTERPLOT subcommand yields the scatterplot for the variables. Note that the variables have been standardized (z scores) and then plotted.

(continued)

consider here only one of the achievement variables, knowledge of body parts. In particular, we consider pretest and posttest data on body parts for a sample of 80 children.

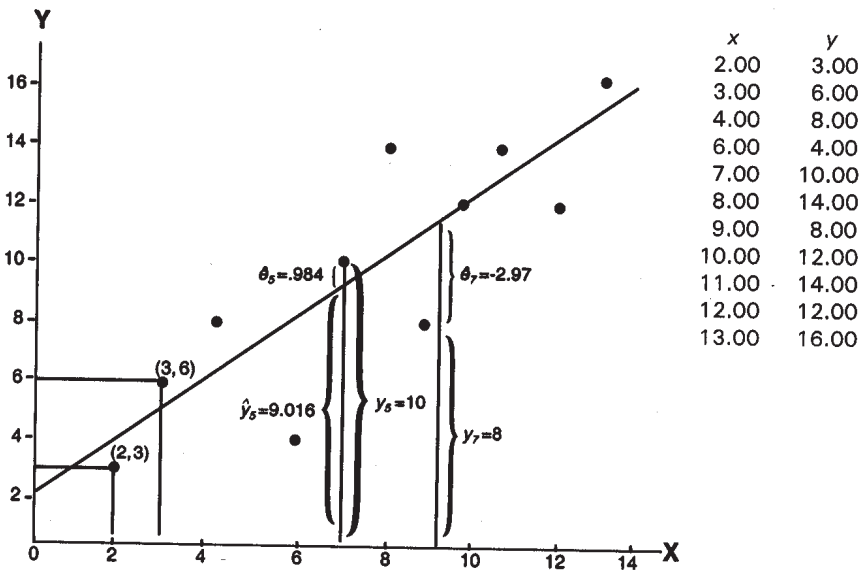


FIGURE 6.2 Errors of Prediction

TABLE 6.2
(Continued)

Placeholder for T0602b from p. 245 of previous edition
--

① This legend means there is one observation whenever a single dot appears, two observations whenever a : appears, and 5 observations where there is an asterisk (*).

② The multiple correlation here is in fact the simple correlation between postbody and prebody, since there is just one predictor.

③ These are the raw coefficients which define the prediction equation: $POSTBODY = .50197 PREBODY + 14.6888$.

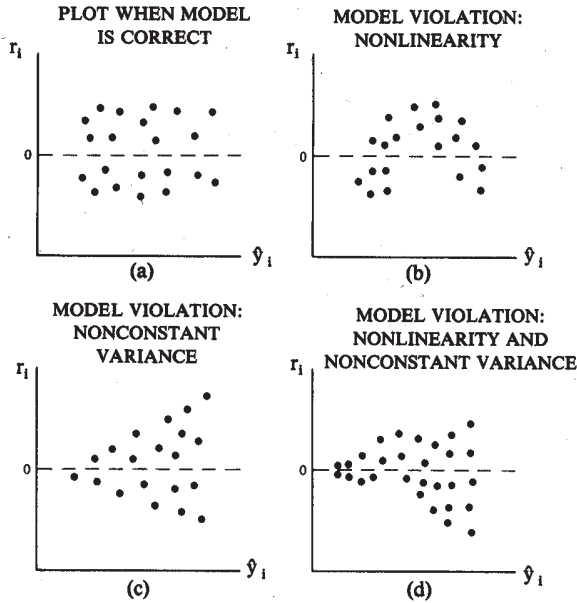


FIGURE 6.3 Plots of Residuals vs. Predicted Values

The command syntax for running the regression analysis, along with selected printout, is presented in Table 6.2. Part of the printout is a standardized scatterplot (recall that when variables are standardized this does *not* affect the magnitude of the correlation).

6.2 ASSUMPTIONS FOR THE ERRORS

The errors (e_i) are assumed to be independent, with constant variance and normally distributed with a mean of 0. If these assumptions are valid for a given set of data, then the estimated errors (\hat{e}_i), called the residuals, should behave similarly. There are various plots involving the residuals that are available for assessing potential problems with a linear regression model. One of the most useful plots, in my opinion, involves graphing the residuals against the predicted values. If the assumptions of the regression model are tenable, then the residuals should scatter randomly about a horizontal line of 0. *Any systematic pattern or clustering of the residuals suggests a model violation(s).* In Figure 6.3 I present four plots: one in which the assumptions are tenable, while in the other 3 plots there is a model violation(s). We obtained this plot for the first data example, and the results are pre-

Placeholder for F0604 from p. 247 of previous edition

FIGURE 6.4 Plots of Residuals vs. Predicted Values for Example 1 Data

sented in Figure 6.4. This plot indicates that the assumptions are tenable for this set of data.

6.3 INFLUENTIAL DATA POINTS

There is one additional point we wish to make before moving into multiple regression. There are situations where a single point may have a big influence on the resulting prediction equation; such a point is called an influential point. A statistic that is quite useful for detecting such influential points is called *Cook's distance*. Cook and Weisberg (1982) indicate that if Cook's distance (which is readily obtained from SPSS or SAS) is > 1 , then generally that point will be influential. As a vivid illustration of such a point, consider Case B from Chapter 1, Section 1.6. The last data point (24,5) is an outlier. Without that point we get a nice regression line. When that point is included, however, it pulls the regression line down considerably. The two regression lines are shown in Figure 6.5.

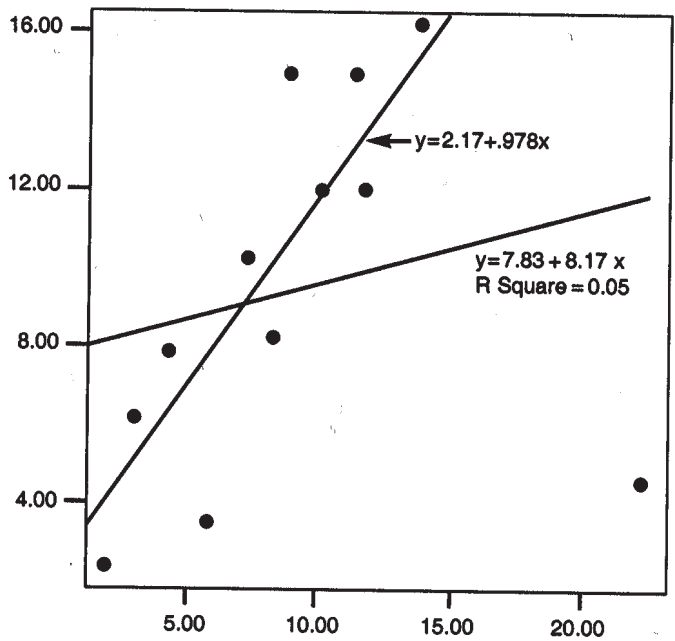


FIGURE 6.5 Regression Lines With and Without Influential Point

How Summary Statistics Can Be Misleading

One of the reviewers of this second edition noted that data sets provided by Anscombe (1973) can show why summary statistics can be misleading, and hence the need for plotting the data. I commented on these data sets in the first edition of my multivariate text (Stevens, 1986, p. 86), but did not show the plots. The actual data are as follows:

X	Y1	Y2	Y3
4	4.26	3.10	5.39
5	5.68	4.74	5.73
6	7.24	6.13	6.08
7	4.82	7.26	6.42
8	6.95	8.14	6.77
9	8.81	8.77	7.11
10	8.04	9.14	7.46
11	8.33	9.26	7.81
12	10.84	9.13	8.15
13	7.58	8.74	12.74
14	9.96	8.10	8.84

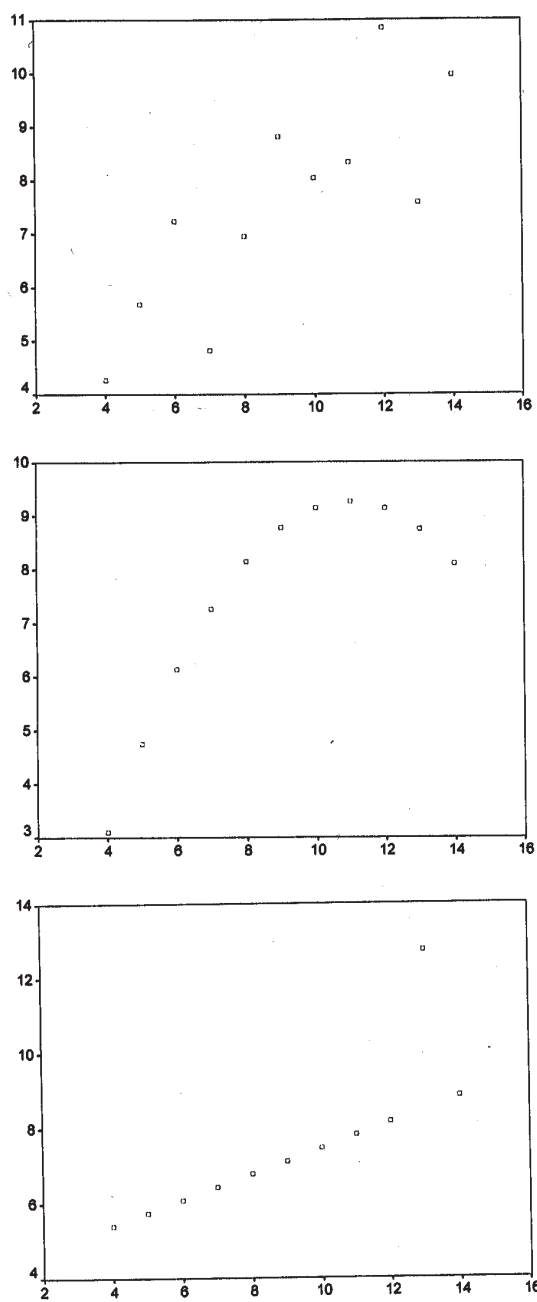


FIGURE 6.6 Plots for Anscombe (1973) Data Sets

These data sets have exactly the same correlation (.816) and the same regression line: $y = 3 + .5x$. Yet the situations are quite different. The plots, from SPSS for Windows 12.0, are given in Figure 6.6. These plots show that only in the first case are the summary statistics an accurate indication of the situation. In the second case there is a curvilinear relationship, and in the last case there is an outlier.

6.4 MULTIPLE REGRESSION

In multiple regression we are interested in predicting a dependent variable from a set of predictors. Since human behavior is complex and influenced by many factors, single predictor studies are limited in their predictive power. For example, in a college GPA study, we are able to predict college GPA better by considering predictors other than high school GPA. Some other factors would be scores on standardized tests (verbal and quantitative), and some noncognitive variables, such as study habits and attitude toward education. That is, we look to other predictors (often test scores) that tap other aspects of criterion behavior.

Consider three other examples of multiple regression studies:

1. Feshbach, Adelman, and Williamson (1977) conducted a study of 850 middle class children. The children were measured in kindergarten on a battery of variables: WPPSI, deHirsch—Jansky Index (assessing various linguistic and perceptual motor skills), the Bender Motor Gestalt, and a Student Rating Scale developed by the authors that measures various cognitive and affective behaviors and skills. These measures were used to predict reading achievement for these same children in grades 1, 2, and 3.
2. Crystal (1988) attempted to predict chief executive officer (CEO) pay for the top 100 of last year's Fortune 500 and the 100 top entries from last year's Service 500. He used the following predictors: company size, company performance, company risk, government regulation, tenure, location, directors, ownership, and age. He found that only about 39% of the variance in CEO pay can be accounted for by these factors.
3. Agresti (1990) gives an example based on real data for 93 homes that were sold in Florida. The dependent variable is price of the home and the predictors were size, number of bathrooms, number of bedrooms, and whether the home was new or not.

In discussing simple regression we mentioned that least squares was used to estimate the parameters and that this procedure minimized the sum of the squared errors of prediction. In multiple regression we will use least squares again. It is very important for the reader to realize that *minimizing the sum of the squared errors of prediction is equivalent to maximizing the correlation between the observed and*

predicted scores. This maximized correlation is called the multiple correlation, i.e., $R = r_{y_i, \hat{y}_i}$. Nunnally (1978) characterized the procedure as “wringing out the last ounce of predictive power” (obtained from the linear combination of the x s, i.e., from the regression equation). Since the correlation is maximum for the sample from which it is derived, when the regression equation is applied to an *independent* sample from the same population (i.e., cross-validated) the predictive power drops off. If the predictive power drops off sharply, then the equation is of very limited utility. That is, it has little generalizability, and hence is of limited scientific value. After all, we derive the prediction equation for the purpose of predicting with it on future (other) samples. If the equation does not predict well on other samples, then it is not fulfilling the purpose for which it was designed.

Sample size (n) and the number of predictors (k) are two crucial factors that determine how well a given equation will cross validate. In particular, the n/k ratio is crucial. For small ratios (5:1 or less) the shrinkage can be substantial.

Since the rest of this chapter is rather lengthy, we give the reader an overview of the critical topics. As we will show shortly, how the predictors are correlated can have a big impact on the multiple correlation. This will take us into the topic of multicollinearity (Section 6.7). In Section 6.8 we discuss several methods for selecting a “good” set of predictors. In Section 6.9 we give two computer examples, using real data, to illustrate some of the methods discussed in Section 6.8. In Section 6.10 we discuss assumptions underlying the regression analysis, and how they can be checked. The crucial topic of model validation is discussed in Section 6.11. Since multiple regression is a *mathematical maximization* procedure, it is very important to check the generalizability of the equation.

6.5 BREAKDOWN OF SUM OF SQUARES IN REGRESSION AND F TEST FOR MULTIPLE CORRELATION

In analysis of variance we broke down variability about the grand mean into between and within variability. In regression analysis variability about the mean is broken down into variability due to regression and variability about the regression. To get at the breakdown, we start with the following identity:

$$y_i - \hat{y}_i = (y_i - \bar{y}) - (\hat{y}_i - \bar{y})$$

Now we square both sides, obtaining

$$(y_i - \hat{y}_i)^2 = [(y_i - \bar{y}) - (\hat{y}_i - \bar{y})]^2$$

Then we sum over the subjects, from 1 to n :

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [(y_i - \bar{y}) - (\hat{y}_i - \bar{y})]^2$$

By algebraic manipulation (see Draper & Smith, 1981, pp. 17–18), this can be rewritten as:

$$\begin{array}{llll} \sum (y_i - \bar{y})^2 & = & \sum (y_i - \hat{y}_i)^2 & + & \sum (\hat{y}_i - \bar{y})^2 \\ \text{sum of squares} & & \text{sum of squares} & & \text{sum of squares} \\ \text{about mean} & = & \text{about regression} & + & \text{due to regression} \\ & & (SS_{res}) & & (SS_{reg}) \\ df : n - 1 & = & (n - k - 1) & + & k (df = \text{degrees of freedom}) \end{array}$$

This results in the following analysis of variance table and the F test for determining whether the population multiple correlation is different from 0.

Source	SS	df	MS	F
Regression	SS_{reg}	k	S_{reg}/k	$\frac{MS_{reg}}{MS_{res}}$
Residual (error)	SS_{res}	$n - k - 1$	$SS_{res}/(n - k - 1)$	MS_{res}

Recall that since the residual for each subject is $\hat{e}_i = y_i - \hat{y}_i$, the mean square error term can be written as $MS_{res} = \sum \hat{e}_i^2 / (n - k - 1)$. Now, R^2 (squared multiple correlation) is given by:

$$R^2 = \frac{\begin{array}{c} \text{sum of squares} \\ \text{due to regression} \end{array}}{\begin{array}{c} \text{sum of squares} \\ \text{about the mean} \end{array}} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{SS_{reg}}{SS_{tot}}$$

Thus, R^2 measures the proportion of total variance on y that is accounted for by the set of predictors. By simple algebra then we can rewrite the F test in terms of R^2 as follows:

$$F = \frac{R^2 / k}{(1 - R^2) / (n - k - 1)} \text{ with } k \text{ and } (n - k - 1) \text{ df} \quad (1)$$

We feel this test is of limited utility, since it does *not necessarily* imply that the equation will cross-validate well, and this is the crucial issue in regression analysis.

Example 3

An investigator obtains $R^2 = .50$ on a sample of 50 subjects with 10 predictors. Do we reject the null hypothesis that the population multiple correlation $= 0$?

$$F = \frac{.50/10}{(1-.50)/(50-10-1)} = 3.9 \text{ with 10 and 39 } df$$

This is significant at .01 level, since the critical value is 2.8.

However, since the n/k ratio is only 5/1, the prediction equation will probably not predict well on other samples and is therefore of questionable utility.

Myers' (1990) response to the question of what constitutes an acceptable value for R^2 is illuminating:

This is a difficult question to answer, and, in truth, what is acceptable depends on the scientific field from which the data were taken. A chemist, charged with doing a linear calibration on a high precision piece of equipment, certainly expects to experience a very high R^2 value (perhaps exceeding .99), while a behavioral scientist, dealing in data reflecting human behavior, may feel fortunate to observe an R^2 as high as .70. An experienced model fitter senses when the value of R^2 is large enough, given the situation confronted. Clearly, some scientific phenomena lend themselves to modeling with considerably more accuracy than others. (p. 37)

His point is that how well one can predict depends on *context*. In the physical sciences, generally quite accurate prediction is possible. In the social sciences, where we are attempting to predict human behavior (which can be influenced by many systematic and some idiosyncratic factors), prediction is much more difficult.

6.6 RELATIONSHIP OF SIMPLE CORRELATIONS TO MULTIPLE CORRELATION

The ideal situation, in terms of obtaining a high R would be to have each of the predictors significantly correlated with the dependent variable and for the predictors to be uncorrelated with each other, so that they measure different constructs and are able to predict different parts of the variance on y . Of course, in practice we will not find this because almost all variables are correlated to some degree. A good situation in practice then would be one in which most of our predictors correlate significantly with y and the predictors have relatively low correlations among themselves. To illustrate the above points further, consider the following three patterns of intercorrelations for three predictors.

		X_1	X_2	X_3			X_1	X_2	X_3			X_1	X_2	X_3
(1)	Y	.20	.10	.30	(2)	Y	.60	.50	.70	(3)	Y	.60	.70	.70
	X_1		.50	.40		X_1		.20	.30		X_1		.70	.60
	X_2			.60		X_2			.20		X_2			.80

In which of these cases would you expect the multiple correlation to be the largest and the smallest respectively? Here it is quite clear that R will be the smallest for 1 because the highest correlation of any of the predictors with y is .30, whereas for the other two patterns at least one of the predictors has a correlation of .70 with y . Thus, we know that R will be at least .70 for cases 2 and 3, whereas for case 1 we only know that R will be at least .30. Furthermore, there is no chance that R for case 1 might become larger than that for cases 2 and 3, because the intercorrelations among the predictors for 1 are approximately as large or larger than those for the other two cases.

We would expect R to be largest for case 2 because each of the predictors is moderately to strongly tied to y and there are low intercorrelations (i.e., little redundancy) among the predictors, exactly the kind of situation we would hope to find in practice. We would expect R to be greater in case 2 than in case 3, because in case 3 there is considerable redundancy among the predictors. Although the correlations of the predictors with y are slightly higher in case 3 (.60, .70, .70) than in case 2 (.60, .50, .70), the much higher intercorrelations among the predictors for case 3 will severely limit the ability of X_2 and X_3 to predict additional variance beyond that of X_1 (and hence significantly increase R), whereas this will not be true for case 2.

6.7 MULTICOLLINEARITY

When there are moderate to high intercorrelations among the predictors, as is the case when several cognitive measures are used as predictors, the problem is referred to as *multicollinearity*. Multicollinearity poses a real problem for the researcher using multiple regression for three reasons:

1. It severely limits the size of R , because the predictors are going after much of the same variance on y . A study by Dizney and Gromen (1967) illustrates very nicely how multicollinearity among the predictors limits the size of R . They studied how well reading proficiency (x_1) and writing proficiency (x_2) would predict course grade in college German. The following correlation matrix resulted:

	x_1	x_2	y
x_1	1.00	.58	.33
x_2		1.00	.45
y			1.00

Note the multicollinearity for x_1 and x_2 ($r_{x_1x_2} = .58$), and also that x_2 has a simple correlation of .45 with y . The multiple correlation R was only .46. Thus, the relatively high correlation between reading and writing severely limited the ability of reading to add hardly anything (only .01) to the prediction of German grade above and beyond that of writing.

2. Multicollinearity makes determining the importance of a given predictor difficult because the effects of the predictors are confounded due to the correlations among them.

3. Multicollinearity increases the variances of the regression coefficients. The greater these variances, the more unstable the prediction equation will be.

The following are two methods for diagnosing multicollinearity:

1. Examine the simple correlations among the predictors from the correlation matrix. These should be observed, and are easy to understand, but the researcher need be warned that they do not always indicate the extent of multicollinearity. More subtle forms of multicollinearity may exist. One such more subtle form is discussed next.
2. Examine the variance inflation factors for the predictors.

The quantity $1/(1-R_j^2)$ is called the j th *variance inflation factor*, where R_j^2 is the squared multiple correlation for predicting the j th predictor from all other predictors.

The variance inflation factor for a predictor indicates whether there is a strong linear association between it and all the remaining predictors. It is distinctly possible for a predictor to have only moderate and/or relatively weak associations with the other predictors in terms of simple correlations, and yet to have a quite high R when regressed on all the other predictors. When is the value for a variance inflation factor large enough to cause concern? Myers (1990) offers the following suggestion: "Though no rule of thumb on numerical values is foolproof, it is generally believed that if any VIF exceeds 10, there is reason for at least some concern; then one should consider variable deletion or an alternative to least squares estimation to combat the problem" (p. 369). The variance inflation factors are easily obtained from SAS REG (cf. Table 6.6).

There are at least three ways of combating multicollinearity. One way is to combine predictors that are highly correlated. For example, if there are three measures relating to a single construct which have intercorrelations of about .80 or larger,

then add them to form a single predictor. The two other ways (factor analysis and ridge regression) are more advanced; see Stevens (1996).

6.8 MODEL SELECTION

There are various methods available for selecting a good set of predictors:

Substantive Knowledge

As Weisberg (1985) noted, “The single most important tool in selecting a subset of variables for use in a model is the analyst’s knowledge of the substantive area under study” (p. 210). It is important for the investigator to be judicious in his/her selection of predictors. Far too many investigators have abused multiple regression by “throwing everything in the hopper,” often merely because the variables are available. Cohen (1990), among others, commented on the indiscriminate use of variables: “I have encountered too many studies with prodigious numbers of dependent variables, or with what seemed to me far too many independent variables, or (heaven help us) both.”

There are several good reasons for generally preferring to work with a small number of predictors: (a) principle of scientific parsimony, (b) reducing the number of predictors improves the n/k ratio, and this helps cross validation prospects, and (c) note the following from Lord and Novick (1968):

Experience in psychology and in many other fields of application has shown that it is seldom worthwhile to include very many predictor variables in a regression equation, for the incremental validity of new variables, after a certain point, is usually very low. This is true because tests tend to overlap in content and consequently the addition of a fifth or sixth test may add little that is new to the battery and still relevant to the criterion. (p. 274)

Or consider the following from Ramsey and Schafer (p. 325):

There are two good reasons for paring down a large number of exploratory variables to a smaller set. The first reason is somewhat philosophical: simplicity is preferable to complexity. Thus, redundant and unnecessary variables should be excluded on principle. The second reason is more concrete: unnecessary terms in the model yield less precise inferences.

Sequential Methods

These are the forward, stepwise and backward selection procedures that are very popular with many researchers. All these procedures involve partialling out process; i.e., they look at the contribution of a predictor with the effects of the other predictors partialled out, or held constant. Many readers may have been exposed in a previous statistics course to the notion of a partial correlation, but a review is nevertheless in order.

The partial correlation between variables 1 and 2 with variable 3 partialled from both 1 and 2 is the correlation with variable 3 held constant, as the reader may recall. The formula for the partial correlation is given by:

$$r_{12.3} = (r_{12} - r_{13}r_{23}) / \sqrt{1 - r_{13}^2} \sqrt{1 - r_{23}^2}$$

Let us put this in the context of multiple regression. Suppose we wish to know what the partial of y (dependent variable) is with predictor 2 with predictor 1 partialled out. The formula would be, following what we have above:

$$r_{y2.1} = (r_{y2} - r_{y1}r_{21}) / \sqrt{1 - r_{y1}^2} \sqrt{1 - r_{21}^2}$$

We apply this formula to show how SPSS obtains the partial correlation of .528 for INTEREST in Table 6.4 under EXCLUDED VARIABLES in the first upcoming computer example. In this example CLARITY (abbreviated as *clr*) entered first, having a correlation of .862 with dependent variable INSTEVAL (abbreviated as *inst*). The correlations below are taken from the correlation matrix, given near the beginning of Table 6.4.

$$r_{\text{inst int.clr}} = .435 - (.862)(.20) / \sqrt{1 - .862^2} \sqrt{1 - .20^2}$$

The correlation between the two predictors is .20, as shown above.

We now give a brief description of the forward, stepwise and backward selection procedures.

FORWARD—The first predictor that has an opportunity to enter the equation is the one with the largest simple correlation with y . If this predictor is significant, then the predictor with the largest partial correlation with y is considered, etc. At some stage a given predictor will not make a significant contribution and the procedure terminates. It is important to remember that with this procedure, once a predictor gets into the equation it stays.

STEPWISE—This is basically a variation on the forward selection procedure.. However, at each stage of the procedure a test is made of the least useful predictor. The importance of each predictor is constantly reassessed. Thus, a predictor that may have been the best entry candidate earlier may now be superfluous.

BACKWARD—The steps are as follows: (a) An equation is computed with ALL the predictors. (b) The partial F is calculated for every predictor, treated as

though it were the last predictor to enter the equation. (c) The smallest partial F value, say F_1 , is compared to a preselected significance, say F_0 . If $F_1 < F_0$, remove that predictor and recomputed the equation with the remaining variables. Reenter stage B.

Use of Mallow's C_p

Before we introduce Mallow's C_p , it is important to consider the consequences of underfitting (important variables are left out of the model) and overfitting (having variables in the model that make essentially no contribution or are marginal). Myers (1990, pp. 178–180) has an excellent discussion on the impact of underfitting and overfitting, and notes that, “A model that is too simple may suffer from biased coefficients and biased prediction, while an overly complicated model can result in large variances, both in the coefficients and in the prediction.”

This measure was introduced by Mallow's (1973) as a criterion for selecting a model. It measures total squared error, and it was recommended by Mallow's to choose the model(s) for which $C_p \approx p$, where $p = k + 1$.

All Possible Regressions

If you wish to follow this route, then the SAS REG procedure should be considered. The number of regressions increases quite sharply as k increases, however, the program will efficiently identify good subsets. Good subsets are those which have the smallest Mallow's value.

Use of one or more of the above methods will often yield a number of models of roughly equal efficacy. As Myers noted (1990), “The successful model builder will eventually understand that with many data sets, several models can be fit that would be of nearly equal effectiveness. Thus, the problem that one deals with is the **selection of one model from a pool of candidate models**” (p.164). One of the problems with the stepwise methods, which are frequently used, is that they have led many researchers to conclude they have found the best model, when in fact there may be some better models and/or several other models that are about as good. As Huberty notes (1989), “And one or more of these subsets may be more interesting or relevant in a substantive sense” (p.46).

As mentioned earlier, Mallows criterion is useful in guarding against both underfitting and overfitting. Another very important criterion that can be used to select from the candidate pool relates to the generalizability of the prediction equation, i.e., validating the equation. Three methods of model validation are discussed in 6.11. Briefly, they are:

1. Data splitting—Randomly split the data, obtain a prediction equation on one part of the random split, and then check its predictive power on the other sample.
2. Use of the PRESS statistic.
3. Obtain an estimate of the average predictive power of the equation on many other samples from the same population, using a formula due to Stein (Herzberg, 1969).

The SPSS application guides comment on overfitting and the use of several models. There is no one test to determine the dimensionality of the best submodel. Some researchers find it tempting to include too many variables in the model, which is called overfitting. Such a model will perform badly when applied to a new sample from the same population (cross validation). Automatic stepwise procedures can not do all the work for you. Use them as a tool to determine roughly the number of predictors needed (for example, you might find 3 to 5 variables). If you try several methods of selection, you may identify candidate predictors that are not included by any method. Ignore them and fit models, say, 3 to 5 variables, selecting alternative subsets from among the better candidates. You may find several subsets that perform equally as well. Then knowledge of the subject matter, how accurately individual variables are measured, and what a variable “communicates” may guide selection of the model to report.

This writer doesn’t disagree with the above comments, however, **he would favor the model which cross validates best. If 2 models cross validate about the same, then I would favor the model which makes most substantive sense.**

6.9 TWO COMPUTER EXAMPLES

To illustrate the use of several of the aforementioned model selection methods, we consider two computer examples. The first example illustrates the SPSS REGRESSION program, and uses data from Morrison (1983) on 32 students enrolled in an MBA course. We predict instructor course evaluation from 5 predictors. The second example illustrates SAS REG on quality ratings of 46 research doctorate programs in psychology, where we are attempting to predict quality ratings from factors such as number of program graduates, percentage of graduates that received fellowships or grant support, etc. (Singer & Willett, 1988).

Example 6—SPSS REGRESSION on Morrison MBA Data

The data for this problem are from Morrison (1983). The dependent variable is instructor course evaluation in an MBA course, with the five predictors being clarity, stimulation, knowledge, interest, and course evaluation. We illustrate two of the

TABLE 6.3
SPSS Control Syntax for Stepwise Regression Run on MORRISON Data
and Correlation Matrix

TITLE 'MULTIPLE REGRESSION-MORRISON DATA'.						
DATA LIST FREE/INSTEVAL CLARITY STIMUL KNOWLEDG INTEREST						
COUEVAL.						
BEGIN DATA.						
1	1	2	1	1	2	1
2	1	3	2	2	2	2
2	2	3	1	3	3	2
2	2	2	1	1	2	2
2	2	2	4	2	2	2
2	3	2	1	1	2	3
3	4	3	2	2	3	3
3	4	5	1	1	3	3
3	3	3	2	1	3	3
1	2	2	1	1	1	1
1	1	1	1	1	1	2
2	3	3	1	1	2	2
2	2	3	2	1	2	2
2	3	3	1	1	3	2
2	3	4	1	1	2	3
3	4	3	1	2	3	3
3	4	3	1	1	2	3
3	4	4	1	2	3	3
4	5	5	2	3	4	4
4	4	5	2	3	4	4
END DATA.						
LIST.						
① REGRESSION DESCRIPTIVES = DEFAULT/						
VARIABLES = INSTEVAL TO COUEVAL/						
② STATISTICS = DEFAULT TOL SELECTION/						
DEPENDENT = INSTEVAL/						
③ METHODS = STEPWISE/						
④ CASEWISE = ALL PRED RESID ZRESID LEVER COOK/						
⑤ SCATTERPLOT (*RES,*PRE)/.						

CORRELATION MATRIX						
	INSTEVAL	CLARITY	STIMUL	KNOWLEDGE	INTEREST	COUEVAL
INSTEVAL	1.000	.862	.739	.282	.435	.738
CLARITY	.862	1.000	.617	.057	.200	.651
STIMUL	.739	.617	1.000	.078	.317	.523
KNOWLEDGE	.282	.057	.078	1.000	.583	.041
INTEREST	.435	.200	.317	.583	1.000	.448
COUEVAL	.738	.651	.523	.041	.448	1.000

① The DESCRIPTIVES = DEFAULT subcommand yields the means, standard deviations and the correlation matrix for the variables.

② This STATISTICS subcommand TOL part yields useful information concerning multicollinearity. In particular it yields the VIF's (variance inflation factors). The SELECTION part yields, among other things, Mallows' prediction criterion, which is very useful in selecting a set of predictors.

③ To obtain the backward selection procedure, we would simply put METHOD = BACK-WARD/

④ This CASEWISE subcommand yields important regression diagnostics: ZRESID (standardized residuals—for identifying outliers on y), LEVER (hat elements—for identifying outliers on predictors), and COOK (Cook's distance—for identifying influential data points).

⑤ This SCATTERPLOT subcommand yields the plot of the residuals vs. the predicted values, which is very useful for determining whether any of the assumptions underlying the linear regression model may be violated.

TABLE 6.4
Selected Printout From SPSS Syntax Editor Stepwise Regression Run
on the Morrison MBA Data

Placeholder for T0604a

From p. 263 of previous edition

TABLE 6.4
(Continued)

Placeholder for T0604b from p. 264 of previous edition

TABLE 6.4
(Continued)

Place holder for T0604c from p. 265 of previous edition

sequential procedures, stepwise and backward selection, using the SPSS REGRESSION program. The control lines for running the analyses, along with the correlation matrix, are given in Table 6.3.

SPSS REGRESSION has “ p values,” denoted by PIN and POUT, which govern whether a predictor will enter the equation and whether it will be deleted. The default values are PIN = .05 and POUT = .10. In other words, a predictor must be “significant” at the .05 level to enter, or must not be significant at the .10 level to be deleted.

First, we discuss the stepwise procedure results. Examination of the correlation matrix in Table 6.3 reveals that three of the predictors (CLARITY, STIMUL, and COUEVAL) are strongly related to INSTEVAL (simple correlations of .862, .739, and .738, respectively). Because clarity has the highest correlation, it will enter the equation first. Superficially, it might appear that STIMUL or COUEVAL would enter next; however, we must take into account how these predictors are correlated with CLARITY, and indeed both have fairly high correlations with CLARITY (.617 and .651 respectively). Thus, they will not account for as much unique variance on INSTEVAL, above and beyond that of CLARITY, as first appeared. On the other hand, INTEREST, which has a considerably lower correlation with INSTEVAL (.44), is only correlated .20 with CLARITY. Thus, the variance on INSTEVAL it accounts for is relatively independent of the variance CLARITY accounted for. And, as seen in Table 6.4, it is INTEREST that enters the regression equation second.

TABLE 6.5
Selected Printout From SPSS Regression for Backward Selection
on the Morrison MBA Data

Placeholder for T0605a from p. 266 of previous edition.

TABLE 6.5
(Continued)

Placeholder for T0605b from p. 267 of previous edition.

STIMUL is the third and final predictor to enter, since its p value (.0086) is less than the default value of .05. Finally, the other predictors (KNOWLEDGE and COUEVAL) don't enter since their p values (.0989 and .1288) are greater than .05.

Selected printout from the backward selection procedure appears in Table 6.5. First, all of the predictors are put into the equation. Then, the procedure determines which of the predictors makes the *least* contribution when entered last in the equation. That predictor is INTEREST, and since its p value is .9097, it is deleted from the equation. None of the other predictors can be further deleted because their p values are much less than .10.

Interestingly, note that two *different* sets of predictors emerge from the two sequential selection procedures. The stepwise procedure yields the set (CLARITY, INTEREST, and STIMUL), while the backward procedure yields the set (COUEVAL, KNOWLEDGE, STIMUL, and CLARITY). However, CLARITY and STIMUL are common to both sets. On the grounds of parsimony, we might prefer the set (CLARITY, INTEREST, and STIMUL), especially since the adjusted R^2 s for the two sets are quite close (.84 and .87).

There are three other things that should be checked out before settling on this as our chosen model:

- 1. We need to determine if the assumptions of the linear regression model are tenable.**
- 2. We need an estimate of the cross-validity power of the equation.**
- 3. We need to check for the existence of outliers and/or influential data points.**

Figure 6.4 showed the plot of the residuals versus the predicted values from SPSS. This plot showed essentially random variation of the points about the horizontal line of 0, indicating no violations of assumptions.

The issues of cross-validity power and outliers are considered later in this chapter, and are applied to this problem in Section 6.15, after both topics have been covered.

TABLE 6.6
SAS REG Control Lines for Stepwise and MAXR Runs on the National
Academy of Sciences Data and the Correlation Matrix

<pre>DATA SINGER; INPUT QUALITY NFACUL NGRADS PCTSUPP PCTGRT NARTIC PCTPUB; CARDS; DATA LINES ① PROC REG SIMPLE CORR; ② MODEL QUALITY = NFACUL NGRADS PCTSUPP PCTGRT NARTIC PCTPUB/ SELECTION = STEPWISE VIF R INFLUENCE; MODEL QUALITY = NFACUL NGRADS PCTSUPP PCTGRT NARTIC PCTPUB/ SELECTION = MAXR VIF R INFLUENCE;</pre>																																																																																									
<p>① SIMPLE is needed to obtain descriptive statistics (means, variances, etc) for all variables. CORR is needed to obtain the correlation matrix for the variables.</p> <p>② In this MODEL statement, the dependent variable goes on the left and all predictors to the right of the equals. SELECTION is where we indicate which of the 9 procedures we wish to use. There is a wide variety of other information we can get printed out. Here we have selected VIF (variance inflation factors), R (analysis of residuals—standard residuals, hat elements, Cooks D), and INFLUENCE (influence diagnostics).</p> <p>Note that there are two separate MODEL statements for the two regression procedures being requested. Although multiple procedures can be obtained in one run, you <i>must</i> have separate MODEL statement for each procedure.</p>																																																																																									
<p>CORRELATION MATRIX</p> <table><tr><td></td><td></td><td>NFACUL</td><td>NGRADS</td><td>PCTSUPP</td><td>PCTGRT</td><td>NARTIC</td><td>PCTPUB</td><td>QUALITY</td></tr><tr><td></td><td>2</td><td>3</td><td>4</td><td>5</td><td>6</td><td>7</td><td></td><td>1</td></tr><tr><td>NFACUL</td><td>2</td><td>1.000</td><td></td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>NGRADS</td><td>3</td><td>0.692</td><td>1.000</td><td></td><td></td><td></td><td></td><td></td></tr><tr><td>PCTSUPP</td><td>4</td><td>0.395</td><td>0.337</td><td>1.000</td><td></td><td></td><td></td><td></td></tr><tr><td>PCTGRT</td><td>5</td><td>0.162</td><td>0.071</td><td>0.351</td><td>1.000</td><td></td><td></td><td></td></tr><tr><td>NARTIC</td><td>6</td><td>0.755</td><td>0.646</td><td>0.366</td><td>0.436</td><td>1.000</td><td></td><td></td></tr><tr><td>PCTPUB</td><td>7</td><td>0.205</td><td>0.171</td><td>0.347</td><td>0.490</td><td>0.593</td><td>1.000</td><td></td></tr><tr><td>QUALITY</td><td>1</td><td>0.622</td><td>0.418</td><td>0.582</td><td>0.700</td><td>0.762</td><td>0.585</td><td>1.000</td></tr></table>											NFACUL	NGRADS	PCTSUPP	PCTGRT	NARTIC	PCTPUB	QUALITY		2	3	4	5	6	7		1	NFACUL	2	1.000							NGRADS	3	0.692	1.000						PCTSUPP	4	0.395	0.337	1.000					PCTGRT	5	0.162	0.071	0.351	1.000				NARTIC	6	0.755	0.646	0.366	0.436	1.000			PCTPUB	7	0.205	0.171	0.347	0.490	0.593	1.000		QUALITY	1	0.622	0.418	0.582	0.700	0.762	0.585	1.000
		NFACUL	NGRADS	PCTSUPP	PCTGRT	NARTIC	PCTPUB	QUALITY																																																																																	
	2	3	4	5	6	7		1																																																																																	
NFACUL	2	1.000																																																																																							
NGRADS	3	0.692	1.000																																																																																						
PCTSUPP	4	0.395	0.337	1.000																																																																																					
PCTGRT	5	0.162	0.071	0.351	1.000																																																																																				
NARTIC	6	0.755	0.646	0.366	0.436	1.000																																																																																			
PCTPUB	7	0.205	0.171	0.347	0.490	0.593	1.000																																																																																		
QUALITY	1	0.622	0.418	0.582	0.700	0.762	0.585	1.000																																																																																	

Example 7—SAS REG on Doctoral Programs
in Psychology

The data for this example come from a National Academy of Sciences report (Jones, Lindzey, & Coggsdall, 1982) that, among other things, provided ratings on the quality of 46 research doctoral programs in psychology. The six variables used to predict quality are:

NFACULTY—number of faculty members in the program as of December 1980.

NGRADS—number of program graduates from 1975 through 1980.

PCTSUPP—percentage of program graduates from 1975–1979 that received fellowships or training grant support during their graduate education.

PCTGRANT—percentage of faculty members holding research grants from the Alcohol, Drug Abuse, and Mental Health Administration, the National Institute of Health or the National Science Foundation at any time during 1978–1980.

NARTICLE—number of published articles attributed to program faculty members from 1978–1980.

PCTPUB—percentage of faculty with one or more published articles from 1978–1980.

Both the stepwise procedure and the MAXR procedure were used on this data to generate several regression models. The control lines for doing this, along with the correlation matrix, are given in Table 6.6.

The stepwise procedure terminated after 4 predictors entered. Below is the summary table, exactly as it appears on the printout:

Summary of Stepwise Procedure for Dependent Variable QUALITY							
Step	Variable Entered	Removed	Partial R^{*2}	Model R^{*2}	$C(p)$	F	$Prob > F$
1	NARTIC		0.5809	0.5809	55.1185	60.9861	0.0001
2	PCTGRT		0.1668	0.7477	18.4760	28.4156	0.0001
3	PCTSUPP		0.0569	0.8045	7.2970	12.2197	0.0011
4	NFACUL		0.0176	0.8221	5.2161	4.0595	0.0505

This four predictor model appears to be a reasonably good one. First, Mallows' C_p is very close to p (recall $p = k + 1$), that is, $5.216 \approx 5$, indicating that there is not much bias in the model. Second, $R^2 = .8221$, indicating that we can predict quality quite well from the 4 predictors. Although this R^2 is *not* adjusted, the adjusted value will not differ much because we have not selected from a large pool of predictors.

Selected printout from the MAXR procedure run appears in Table 6.7. From Table 6.7 we can construct the following results:

BEST MODEL	VARIABLE(S)	MALLOWS' C_p
for 1 variable	NARTIC	55.118
for 2 variables	PCTGRT, NFACUL	16.859
for 3 variables	PCTPUB, PCTGRT, NFACUL	9.147
for 4 variables	NFACUL, PCTSUPP, PCTGRT, NARTIC	5.216

In this case, the *same* 4 predictor model is selected by the MAXR procedure that was selected by the stepwise procedure.

TABLE 6.7
Selected Printout From the MAXR Run on the National Academy of Sciences Data

Placeholder for T0607 from p. 269 of previous edition.

Caveat on p Values for the “Significance” of Predictors

The p values that are given by SPSS and SAS for the “significance” of each predictor at each step for stepwise or the forward selection procedures should be treated tenuously, especially if your initial pool of predictors is moderate (15) or large (30). The reason is that the ordinary F distribution is *not* appropriate here, because the largest F is being selected out of all F s available. Thus, the appropriate critical value will be larger (and can be considerably larger) than would be obtained from the ordinary null F distribution. Draper and Smith (1981) note, “Studies have shown, for example, that in some cases where an entry F test was made at the α level, the appropriate probability was $q\alpha$, where there were q entry candidates at that stage” (p. 311). This is saying, for example, that an experimenter may think his or her probability of erroneously including a predictor is .05, when in fact the *actual* probability of erroneously including the predictor is .50 (if there were 10 entry candidates at that point)!

Thus, the F tests are positively biased, and the greater the number of predictors, the larger the bias. Hence, these F tests should be used only as rough guides to the usefulness of the predictors chosen. The acid test is how well the predictors do under cross-validation. It can be unwise to use *any* of the stepwise procedures with 20 or 30 predictors and only 100 subjects, since capitalization on chance is great, and the results may well not cross-validate. To find an equation that probably will have generalizability, it is best to carefully select (using substantive knowledge and/or any previous related literature) a small or relatively small set of predictors.

6.10 CHECKING ASSUMPTIONS FOR THE REGRESSION MODEL

Recall that in the linear regression model it is assumed that the errors are independent and follow a normal distribution with constant variance. The normality assumption can be checked through use of the histogram of the standardized residuals. The independence assumption implies that the subjects are responding independently of one another. This is an important assumption. Even a slight violation can cause the type I error rate to be several times greater than what one desires.

Let us consider a situation where the independence assumption would not be tenable. Suppose we had 50 college freshmen each write 4 in class essays. Then, although we have 200 essays to grade, we have only 50 independent responses, since the responses for each student are going to be correlated.

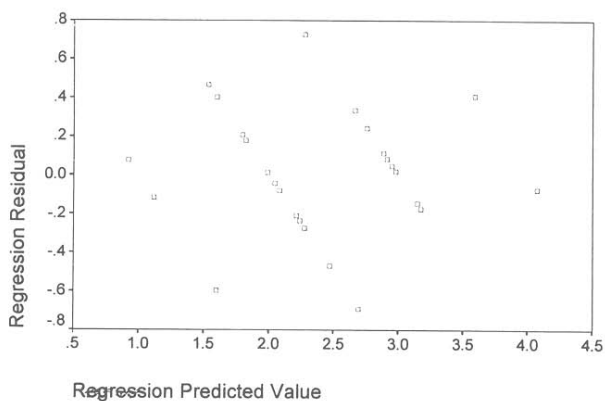
Residual Plots

There are various plots available for assessing potential problems with the regression model (Draper & Smith, 1981; Weisberg, 1985). A very useful plot graphs the standardized residuals (y) vs. the predicted values (x). If the assumptions of the linear regression model are tenable, then the standardized residuals should scatter randomly about a horizontal line of 0, as shown in Figure 6.3a (see Section 6.3). *Any systematic pattern or clustering of the residuals suggests a model violation(s).* Three such systematic patterns are shown in Figures 6.3b to 6.3d. Figure 6.3b shows a systematic quadratic (second degree equation) clustering of the residuals. For Figure 6.3c the variability of the residuals increases systematically as the predicted values increase, suggesting a violation of the constant variance assumption.

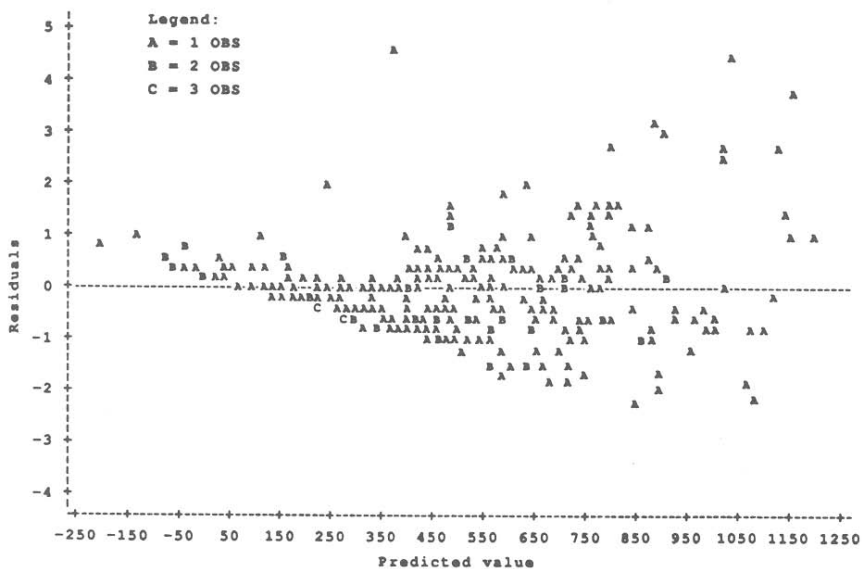
In Figure 6.7 we present residual plots for three real data sets. The first plot is for the Morrison data (the first computer example), and shows essentially random scatter of the residuals, suggesting no violations of assumptions. The remaining two plots are from a study by a statistician who analyzed the salaries of over 260 major league hitters, using predictors such as career batting average, career home runs per time at bat, years in the major leagues, etc. These plots are from Moore and McCabe (1989), and are used with permission. Figure 6.7b, which plots the residuals versus predicted salaries, shows a clear violation of the constant variance assumption. For lower predicted salaries there is little variability about 0, but for the high salaries there is considerable variability of the residuals. The implication of this is that the model will predict lower salaries quite accurately, but not so for the higher salaries.

Figure 6.7c plots the residuals versus number of years in the major leagues. This plot shows a clear curvilinear clustering, that is, quadratic. The curved lines encompass the vast majority of points to make this trend even more evident. The implication of this curvilinear trend is that the regression model will tend to overestimate the salaries of players who have been in the major leagues only a few years or over 15 years, while it will underestimate the salaries of players who have been in the majors about 5 to 9 years.

In concluding this section, note that if nonlinearity or nonconstant variance is found, there are various remedies. For nonlinearity, perhaps a polynomial model is needed. Or sometimes a transformation of the data will enable a nonlinear model to be approximated by a linear one. For nonconstant variance, weighted least squares is one possibility, or more commonly, a variance stabilizing transformation (such as square root or log) may be used. I refer the reader to Weisberg (1985, Chapter 6) for an excellent discussion of remedies for regression model violations.



(a) No Violation



(b) Model Violation: Heterogeneous Variance

FIGURE 6.7 Residual Plots for Three Real Data Sets Showing No Violations, Heterogenous Variance, and Curvilinearity

(Continued)

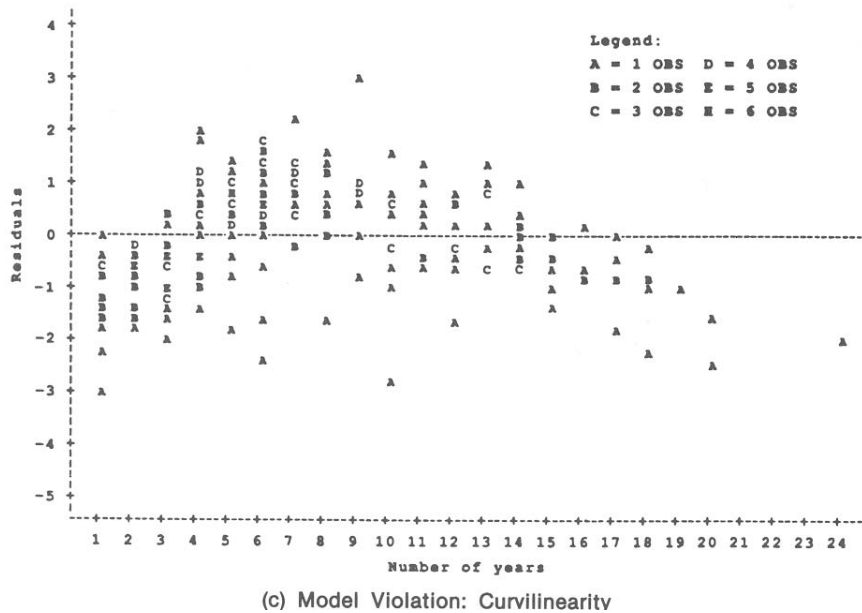


FIGURE 6.7 (Continued)

6.11 MODEL VALIDATION

We indicated earlier that it was crucial for the researcher to obtain some measure of how well the regression equation will predict on an **independent** sample(s) of data. That is, it was important to determine whether the equation had generalizability. We discuss here two methods of model validation: one empirical, and the other involving an estimate of average predictive power on other samples. A third method of model validation, particularly useful when one has a small or moderate sample, utilizes what is called the PRESS statistic. This is a nice empirical measure, but it is more complicated, so I have put it in an Appendix to this chapter for those who are interested. Let me give a brief description of the two methods, and then I will elaborate on each form of validation.

Data Splitting. Here the sample is randomly split in half. It does not have to be split evenly, but we use this for illustration. The regression equation is found on the so-called derivation sample (also called the screening sample, or the sample that “gave birth” to the prediction equation by Tukey). This prediction equation is then applied to the other sample of data (called the validation sample) to see how well it predicts the y score there.

Compute an Adjusted R^2 . There are various adjusted R^2 measures, or measures of shrinkage in predictive power, but they do not estimate the same thing. The one most commonly used, and that which is printed out by SPSS and SAS, is due to Wherry. It is very important to note that the Wherry formula estimates how much variance on y would be accounted for if we had derived the equation in the population from which the sample was drawn. The Wherry formula does *not* indicate how well the derived equation will predict on other samples from the same population. A formula due to Stein (1960) does estimate average cross-validation predictive power. Unfortunately, it was not printed out by SPSS and SAS about 10 years ago, and it is still not printed out by either package.

Data Splitting

Recall that the sample is randomly split. The regression equation is found on the derivation sample and then is applied to the other sample (validation) to determine how well it will predict y there. Below we give a hypothetical example, randomly splitting 100 subjects.

Derivation Sample $n = 50$	Validation Sample $n = 50$		
Prediction Equation $\hat{y}_i = 4 + .3x_1 + .7x_2$	y	x_1	x_2
	6	1	.5
	4.5	2	.3
		
	7	5	.2

Now, using the above prediction equation we predict the y scores in the validation sample:

$$\begin{aligned}\hat{y}_1 &= 4 + .3(1) + .7(.5) = 4.65 \\ \hat{y}_2 &= 4 + .3(2) + .7(.3) = 4.81 \\ &\vdots \\ \hat{y}_{50} &= 4 + .3(5) + .7(.2) = 5.64\end{aligned}$$

The cross-validated R then is the correlation for the following set of scores:

y	\hat{y}_i
6	4.65
4.5	4.81
.	
.	
7	5.64

Adjusted R^2

Herzberg (1969) presents a discussion of various formulas that have been used to estimate the amount of shrinkage found in R^2 . As mentioned earlier, the one most commonly used, and due to Wherry, is given by

$$\hat{\rho}^2 = 1 - \frac{(n-1)}{(n-k-1)}(1-R^2) \quad (2)$$

where $\hat{\rho}$ is the estimate of ρ , the population multiple correlation coefficient. This is the adjusted R^2 printed out by SAS and SPSS. Draper and Smith (1981) comment on Equation 5: “A related statistic...is the so called adjusted r (R_a^2), the idea being that the statistic R_a^2 can be used to compare equations fitted not only to a specific set of data but also to two or more entirely different sets of data. The value of this statistic for the latter purpose is, in our opinion, not high” (p. 92).

Herzberg notes that, “In applications, the population regression function can never be known and one is more interested in how effective the *sample* regression function is in *other* samples. A measure of this effectiveness is r_c , the sample cross-validity. For any given regression function r_c will vary from validation sample to validation sample. The average value of r_c will be approximately equal to the correlation, in the *population*, of the sample regression function with the criterion. This correlation is the population cross-validity, ρ_c . Wherry’s formula estimates ρ rather than ρ_c ” (p. 4).

There are two possible models for the predictors: (1) regression—the values of the predictors are fixed, i.e., we study y only for certain values of x , and (2) correlation—the predictors are random variables—this is a much more reasonable model for social science research. Herzberg presents the following formula for estimating ρ_c^2 under the correlation model:

$$\hat{\rho}_c^2 = 1 - \left(\frac{n-1}{n-k-1} \right) \left(\frac{n-2}{n-k-2} \right) \left(\frac{n+1}{n} \right) (1-R^2) \quad (3)$$

where n is sample size and k is the number of predictors. It can be shown that $\rho_c < \rho$.

If you are interested in cross-validity predictive power, then the Stein formula (Equation 6) should be used. As an example, suppose $n = 50$, $k = 10$, and $R^2 .50$. If you use the Wherry formula (Equation 5), then your estimate is

$$\hat{\rho}^2 = 1 - 49/39(.50) = .372$$

whereas with the proper Stein formula you would obtain

$$\hat{\rho}_c^2 = 1 - (49/39)(48/38)(51/50)(.50) = .191$$

In other words, use of the Wherry formula would give a misleadingly positive impression of the cross validity predictive power of the equation.

Table 6.8 shows how the estimated predictive power drops off using the Stein formula (Equation 6) for small to fairly large subject/variable ratios when $R^2 = .50$.

6.12 IMPORTANCE OF THE ORDER OF THE PREDICTORS IN REGRESSION ANALYSIS

The order in which the predictors enter a regression equation can make a great deal of difference with respect to how much variance on y they account for, especially for moderate or highly correlated predictors. Only for uncorrelated predictors (which would rarely occur in practice) does the order not make a difference. We give two examples to illustrate.

Example 8

A dissertation by Crowder (1975) attempted to predict ratings of trainably mentally retarded individuals (TMs) using I.Q. (x_2) and scores from a TEST of Social Inference (TSI). He was especially interested in showing that the TSI had incremental predictive validity. The criterion was the average ratings by two individuals in charge of the TMs. The intercorrelations among the variables were:

$$r_{x_1x_2} = .59, r_{yx_2} = .54, r_{yx_1} = .566$$

Now, consider two orderings for the predictors, one where TSI is entered first, and the other ordering where I.Q. is entered first.

	First Ordering % of variance		Second Ordering % of variance
TSI	32.04	I.Q.	29.16
I.Q.	6.52	TSI	9.40

The first ordering conveys an overly optimistic view of the utility of the TSI scale. Since we know that I.Q. will predict ratings it should be entered first in the equation (as a control variable), and then TSI to see what its incremental validity is, i.e., how much it *adds* to predicting ratings above and beyond what I.Q. does. Because of the moderate correlation between I.Q. and TSI, the amount of variance accounted for by TSI differs considerably when entered first vs. second (32.04 vs. 9.4).

TABLE 6.8
Estimated Cross Validity Predictive Power for Stein Formula
Stein Estimate Formula

$$\hat{\rho}_e^2 = 1 - \left(\frac{n-1}{n-k-1} \right) \left(\frac{n-2}{n-k-2} \right) \left(\frac{n+1}{n} \right) (1 - R^2) \quad \textcircled{1}$$

Subject/Variable Ratio	Stein Estimate
<i>Small (5:1)</i>	
$N = 50, k = 10, R^2 = .50$.191 $\textcircled{2}$
$N = 50, k = 10, R^2 = .75$.595
$N = 50, k = 10, R^2 = .85$.757
<i>Moderate (10:1)</i>	
$N = 100, k = 10, R^2 = .50$.374
$N = 100, k = 10, R^2 = .75$.690
<i>Fairly Large (15:1)</i>	
$N = 150, k = 10, R^2 = .50$.421

①If there is selection of predictors from a larger set, then the *median* should be used as the *k*. For example, if 4 predictors were selected from a set of 30 predictors by, say, the stepwise procedure, then the median between 4 and 30 (that is, 17) should be the *k* used in the Stein formula.

The 9.4% of variance accounted for by TSI when entered second is obtained through the use of the semipartial correlation previously introduced:

$$r_{y1.2(s)} = \frac{.566 - .54(.59)}{\sqrt{1 - .59^2}} = .306 \Rightarrow r_{y1.2(s)}^2 = 0.94$$

Example 9

Consider the following matrix of correlations for a three predictor problem:

	x_1	x_2	x_3
y	.60	.70	.70
x_1		.70	.60
x_2			.80

How much variance on *y* will x_3 account for if entered first, and if entered last? If x_3 is entered first, then it will account for $(.7)^2 \times 100$ or 49% of the variance on *y*. If x_3 is entered last, we need to compute a second order semipartial correlation (see Stevens, 1996, p. 102 for details). The answer is only 4.8% of the variance on *y*. Because the predictors are so highly correlated, most of the variance on *y* that x_3 could have accounted for has already been accounted for by x_1 and x_2 .

Controlling the Order of Predictors in the Equation

With the forward and stepwise selection procedures, the order of entry of predictors into the regression equation is determined via a mathematical maximization procedure. That is, the first predictor to enter is the one with the largest (maximized) correlation with y , the second to enter is the predictor with the largest partial correlation, etc. However, there are situations where one may not want the mathematics to determine the order of entry of predictors. For example, suppose we have a five predictor problem, with two proven predictors from previous research. The other three predictors are included to see if they have any incremental validity. In this case we would want to enter the two proven predictors in the equation first (as control variables), and then let the remaining three predictors “fight it out” to determine whether any of them add anything significant to predicting y above and beyond the proven predictors.

With SPSS REGRESSION or SAS REG we can control the order of predictors, and in particular, we can *force* predictors into the equation. In Table 6.9 we illustrate how this is done for SPSS and SAS for the above five predictor situation.

6.13 OTHER IMPORTANT ISSUES

Preselection of Predictors

An industrial psychologist hears about the predictive power of multiple regression and is excited. He wants to predict success on the job, and gathers data for 20 potential predictors on 70 subjects. He obtains the correlation matrix for the variables, and then picks out 6 predictors that correlate significantly with success on the job and that have low intercorrelations among themselves. The analysis is run, and the R^2 is highly significant. Furthermore, he is able to explain 52% of the variance on y (more than other investigators have been able to do). Are these results generalizable? Probably not, since what he did involves a *double* capitalization on chance:

1. First, in preselecting the predictors from a larger set, he is capitalizing on chance. Some of these variables would have high correlations with y because of sampling error, and consequently their correlations would tend to be lower in another sample.
2. Second, the mathematical maximization involved in obtaining the multiple correlation involves capitalizing on chance.

Preselection of predictors is common among many researchers, who are unaware of the fact that this tends to make their results sample specific. Nunnally

(1978) has a nice discussion of the preselection problem, and Wilkinson (1979) has shown the considerable positive bias preselection can have on the test of significance of R^2 in forward selection. The following example from his tables illustrates. The critical value for a 4 predictor problem ($n = 35$) at .05 level is .26, while the appropriate critical value for the *same* n and α level, when preselecting 4 predictors from a set of 20 predictors is .51! Unawareness of the positive bias has led to many results in the literature that are not replicable, for as Wilkinson notes, “A computer assisted search for articles in psychology using stepwise regression from 1969 to 1977 located 71 articles. Out of these articles, 66 forward selections analyses reported as significant by the usual F tests were found. Of these 66 analyses, 19 were *not* significant by [his] Table 1.”

It is important to note that both the Wherry and Herzberg formulas do *not* take into account preselection. Hence, the following from Cohen and Cohen (1983) should be seriously considered: “A more realistic estimate of the shrinkage is obtained by substituting for k the *total* number of predictors from which the selection was made” (p. 107). In other words, they are saying if 4 predictors were selected out of 15, use $k = 15$ in the Herzberg formula. While this may be conservative, using 4 will certainly lead to a positive bias. Probably a median value between 4 and 15 would be closer to the mark, although this needs further investigation.

Positive Bias of R^2

A study by Schutz (1977) on California principals and superintendents illustrates how capitalization on chance in multiple regression (if the researcher is unaware of it) can lead to misleading conclusions. Schutz was interested in validating a “contingency theory of leadership,” that is, that success in administering schools calls for different personality styles depending on the social setting of the school. The theory seems plausible, and in what follows we are not criticizing the theory per se, but the empirical validation of it. Schutz’s procedure for validating the theory involved establishing a relationship between various personality attributes (24 predictors) and several measures of administrative success in heterogeneous samples with respect to social setting using multiple regression, that is, find the multiple R for each measure of success on 24 predictors. Then he showed that the magnitude of the relationships was greater for subsamples homogeneous with respect to social setting. The problem was that he had nowhere near adequate sample size for a reliable prediction equation. Below we present the total sample sizes and the subsamples homogeneous with respect to social setting:

	Superintendents	Principals
Total	$n = 77$	$n = 147$
Subsample(s)	$n = 29$	$n_1 = 35, n_2 = 61, n_3 = 36$

TABLE 6.9
Controlling the Order of Predictors and Forcing Predictors Into the
Equation With SPSS REGRESSION and SAS REG

```

SPSS REGRESSION
TITLE 'FORCING X3 AND X4 & USING STEPWISE SELECTION FOR OTHERS' .
DATA LIST FREE/Y X1 X2 X3 X4 X5
BEGIN DATA.
    DATA LINES
END DATA.
REGRESSION VARIABLES = Y X1 X2 X3 X4 X5/
DEPENDENT = Y
① ENTER X3/ENTER X4/STEPWISE/ .
                                SAS REG

DATA FORCEPR;
INPUT Y X1 X2 X3 X4 X5;
CARDS;
    DATA LINES
PROC REG SIMPLE CORR;
② MODEL Y = X3 X4 X1 X2 X5/INCLUDE = 2 SELECTION = STEPWISE;

```

① These two ENTER subcommands will force the predictors in the specific order indicated. Then the STEPWISE subcommand will determine whether any of the remaining predictors (X1, X2, or X5) have semipartial correlations large enough to be “significant.” If we wished to force in predictors X1, X3, and X4 and then use STEPWISE, the subcommand is ENTER X1 X3 X4/STEPWISE/

② The INCLUDE = 2 forces the *first 2* predictors listed in the MODEL statement into the prediction equation. Thus, if we wish to force X3 and X4 we must list them first on the MODEL statement.

Indeed, Schutz did find that the R 's in the homogeneous subsamples were on the average .34 greater than in the total samples; however, this was an artifact of the multiple regression procedure in this case. As Schutz went from total to his subsamples the number of predictors (k) approached sample size (n). For this situation the multiple correlation increases to 1 *regardless* of whether there is any relationship between y and the set of predictors. And in 3 of 4 of Schutz's subsamples the n/k ratios became dangerously close to 1. In particular it is the case that $E(R^2) = k/(n-1)$, when the population multiple correlation = 0 (Morrison, 1976).

To dramatize this, consider subsample 1 for the principals. Then $E(R^2) = 24/34 = .706$, even when there is *no* relationship between y and the set of predictors. The critical value required just for statistical significance of R at .05 is 2.74, which implies $R^2 > .868$, just to be confident that the population multiple correlation is different from 0!

6.14 OUTLIERS AND INFLUENTIAL DATA POINTS

Since multiple regression is a mathematical maximization procedure, it can be very sensitive to data points that “split off” or are different from the rest of the points, that is, to outliers. Just 1 or 2 such points can affect the interpretation of results, and it is certainly moot as to whether 1 or 2 points should be permitted to have such a profound influence. Therefore, it is important to be able to detect outliers and influential points. There is a distinction between the two because a point that is an outlier (either on y or for the predictors) will *not necessarily* be influential in affecting the regression equation.

There are two basic approaches that can be used in dealing with outliers and influential points. We consider the approach of having an arsenal of tools for isolating these important points for further study, with the possibility of deleting some or all of the points from the analysis. The other approach is to develop procedures that are relatively insensitive to wild points (i.e., robust regression techniques).

Data Editing

Outliers and influential cases can occur because of recording errors. Consequently, researchers should give more consideration to the data editing phase of the data analysis process (i.e., *always* listing the data and examining the list for possible errors). There are many possible sources of error, from the initial data collection to the final keypunching. First, some of the data may have been recorded incorrectly. Second, even if recorded correctly, when all of the data are transferred to a single sheet or a few sheets in preparation for keypunching, errors may be made. Finally, even if no errors are made in these first two steps, an error(s) could be made in entering the data into the terminal.

There are various statistics for identifying outliers on y and on the set of predictors, as well as for identifying influential data points. We discuss first, in brief form, a statistic for each, with advice on how to interpret that statistic. Equations for the statistics are given in my multivariate text (Stevens, 1996), along with a more extensive and somewhat technical discussion for those who are interested.

Measuring Outliers on y

For finding subjects whose predicted scores are quite different from their actual y scores (i.e., they do not fit the model well), the *standardized residuals* (r_i) can be used. If the model is correct, then they have a normal distribution with a mean of 0 and a standard deviation of 1. Thus, about 95% of the r_i should lie within two standard deviations of the mean and about 99% within three standard deviations. Therefore, any standardized residual greater than about 3 in absolute value is unusual and should be carefully examined.

Measuring Outliers on Set of Predictors

The *hat elements* (h_{ii}) can be used here. It can be shown that the hat elements lie between 0 and 1, and that the average hat element is p/n , where $p = k + 1$. Because of this, Hoaglin and Welsch (1978) suggest that $2p/n$ may be considered large. However, this can lead to more points than we really would want to examine, and the reader should consider using $3p/n$. For example, with 6 predictors and 100 subjects, any hat element (also called leverage) greater than $3(7)/100 = .21$ should be carefully examined. This is a very simple and useful rule of thumb for quickly identifying subjects who are very different from the rest of the sample on the set of predictors.

Measuring Influential Data Points

An influential data point is one that when deleted produces a substantial change in at least one of the regression coefficients. That is, the prediction equations with and without the influential point are quite different. *Cook's distance* (1977) is very useful for identifying influential points. It measures the *combined* influence of the case being an outlier on y and on the set of predictors. Cook and Weisberg (1982) have indicated that a Cook's distance > 1 *would generally be considered large*. This provides a "red flag," when examining computer printout, for identifying influential points.

All of the above diagnostic measures are easily obtained from SPSS REGRESSION (cf. Table 6.3) or SAS REG (cf. Table 6.6).

6.15 FURTHER DISCUSSION OF THE TWO COMPUTER EXAMPLES

Morrison Data

Recall that for Morrison data the stepwise procedure yielded the more parsimonious model involving 3 predictors: CLARITY, INTEREST, and STIMUL. If we were interested in an estimate of the predictive power in the population, then the Wherry estimate given by Equation 5 is appropriate. This is given in Table 6.4 under model 3 as ADJUSTED R SQUARE .840. Here the estimate is used in a descriptive sense; to describe the relationship in the population. However, if we are interested in the cross-validity predictive power, then the Stein estimate (Equation 6) should be used. The Stein adjusted R^2 in this case is

$$p_c^2 = 1 - (31/38)(30/27)(33/32)(1 - .856) = .82$$

This estimates that if we were to cross-validate the prediction equation on many other samples from the same population, then *on the average* we would account for about 82% of the variance on the dependent variable. In this instance the estimated dropoff in predictive power is very little from the maximized value of 85.56%. The reason is that the association between the dependent variable and the set of predictors is *very* strong. Thus, we can have confidence in the future predictive power of the equation.

It is also important to examine the regression diagnostics to check for any outliers and/or influential data points. Table 6.10 presents the appropriate statistics, as discussed in Section 6.14, for identifying outliers on the dependent variable (standardized residuals), outliers on the set of predictors (hat elements), and influential data points (Cook's distance).

First, we would expect only about 5% of the standardized residuals to be $> |2|$ if the linear model is appropriate. From Table 6.10 we see that 2 of the ZRESID are $> |2|$, and we would expect about $32(.05) = 1.6$, so nothing seems to be awry here. Next, we check for outliers on the set of predictors. The rough "critical value" here is $3p/n = 3(4)/32 = .375$. Since there are no values under LEVER in Table 6.10 exceeding this value, we have no outliers on the set of predictors. Finally, and perhaps most importantly, we check for the existence of influential data points using Cook's D . Recall that Cook (1982) has suggested if $D > 1$, then the point is influential. All the Cook D s in Table 6.10 are far less than 1, so we have no influential data points.

In summary then, the linear regression model is quite appropriate for the Morrison data. The estimated cross validity power is excellent, and there are no outliers or influential data points.

National Academy of Sciences Data

Recall that both the stepwise procedure and the MAXR procedure yielded the same "best" 4-predictor set: NFACUL, PCTSUPP, PCTGRT, and NARTIC. The maximized $R^2 = .8221$, indicating that 82.21% of the variance in quality can be accounted for by these 4 predictors in *this* sample. Now we obtain two measures of the cross-validity power of the equation. First, from the SAS REG printout, we have PREDICTED RESID SS (PRESS) = 1350.33. Furthermore, the variance for QUALITY is 101.438, so that $\Sigma(Y_i - \bar{Y})^2 = 4564.71$. From these numbers we can compute

$$R_{\text{PRESS}}^2 = 1 - (1350.33) / 4564.71 = .7042$$

This is a good measure of the external predictive power of the equation, where we have n validations, each based on $(n-1)$ observations.

The Stein *estimate* of how much variance on the average we would account for if the equation were applied to many other samples is

$$\rho_c^2 = 1 - (45/41)(44/40)(47/46)(1 - .822) = .7804$$

Now we turn to the regression diagnostics from SAS REG, which are presented in Table 6.11. In terms of the standardized residuals for y , there are two that stand out (-3.0154 and 2.5276 for observations 25 and 44). These are for the University of Michigan and Virginia Polytech. In terms of outliers on the set of predictors, using $2p/n = 2(5)/46 = .217$, there are outliers for observation 15 (University of Georgia), observation 25 (University of Michigan again), and observation 30 (North-eastern).

Using the criterion of Cook $D > 1$, there is one influential data point, observation 25 (University of Michigan). Recall that whether a point will be influential is a *joint* function of being an outlier on y and on the set of predictors. In this case, the University of Michigan definitely doesn't fit the model and it differs dramatically from the other psychology departments on the set of predictors. A check of the DFBETAS reveals that it is very different in terms of number of faculty (DFBETA $= -2.7653$), and a scan of the raw data shows the number of faculty at 111, while the average number of faculty members for all the departments is only 29.5. The question needs to be raised as to whether the University of Michigan is "counting" faculty members in a different way from the rest of the schools. For example, are they including part time and adjunct faculty, and if so, is the number of these quite large?

For comparison purposes, the analysis was also run with the University of Michigan deleted. Interestingly, the same 4 predictors emerge from the stepwise procedure, although the results are better in some ways. For example, Mallows' C_p is now 4.5248, whereas for the full data set it was 5.216. Also, the PRESS residual sum of squares is now only 899.92, whereas for the full data set it was 1350.33.

6.16 SAMPLE SIZE DETERMINATION FOR A RELIABLE PREDICTION EQUATION

The reader may recall that in power analysis one is interested in determining a priori how many subjects are needed per group to have, say, power = .80 at the .05 level. Thus, planning is done ahead of time to ensure that one has a good chance of detecting an effect of a given magnitude. Now, in multiple regression the focus is different and the concern, or at least one very important concern, is development of a prediction equation that has generalizability. A study by Park and Dudycha (1974) provides several tables that, given certain input parameters, enable one to determine how many subjects will be needed for a reliable prediction equation.

TABLE 6.10
Regression Diagnostics (Standardized Residuals, Hat Elements,
and Cook's Distance) for Morrison MBA Data

Placeholder for T0610 from p. 284 of previous edition

① These are the predicted values.

② These are the raw residuals, that is, $\hat{e}_i = y_i - \hat{y}_i$. Thus, for the first subject we have $\hat{e}_1 = 1 - 1.1156 = -.1156$.

③ These are the standardized residuals.

④ The hat elements—they have been called leverage elements elsewhere; hence the abbreviation LEVER.

They considered from 3 to 25 random variable predictors, and found that with about 15 subjects per predictor the amount of shrinkage is small ($<.05$) with high probability (.90), if the squared population multiple correlation (ρ^2) is .50. In Table 6.12 we present selected results from the Park and Dudycha study for 3, 4, 8, and 15 predictors.

To use Table 6.12 we need an estimate of ρ^2 , that is, the squared *population* multiple correlation. Unless an investigator has a good estimate from a previous study that used similar subjects and predictors, we feel taking $\rho^2 = .50$ is a reasonable guess for social science research. In the physical sciences, estimates $>.75$ are quite reasonable. If we set $\rho^2 = .50$ and want the loss in predictive power to be less than .05 with probability = .90, then the required sample sizes are as follows:

		Number of Predictors			
$\rho^2 = .50$	$\epsilon = .05$	3	4	8	15
n		50	66	124	214
n/k ratio		16.7	16.7	15.5	14.3

The n/k ratios in all 4 cases are around 15/1.

We had indicated earlier that *generally* about 15 subjects per predictor are needed for a reliable regression equation in the social sciences, that is, an equation that will cross-validate well. There are three converging lines of evidence that support this conclusion:

1. The Stein formula for estimated shrinkage (Table 6.8).
2. My own experience.
3. The results just presented from the Park and Dudycha study.

However, the Park and Dudycha study (cf. Table 6.12) clearly shows that the magnitude of ρ (population multiple correlation) strongly affects how many subjects will be needed for a reliable regression equation. For example, if $\rho^2 = .75$, then for 3 predictors only 28 subjects are needed, whereas 50 subjects were needed for the same case when $\rho^2 = .50$.

Also, from the Stein formula (Table 6.8), you will see if you plug in .40 for R^2 that more than 15 subjects per predictor will be needed to keep the shrinkage fairly small, while if you insert .70 for R^2 , significantly less than 15 will be needed.

6.17 ANOVA AS A SPECIAL CASE OF REGRESSION ANALYSIS

This section is presented to show that ANOVA is just a special case of regression analysis, i.e., the general linear model. Cohen's (1968) seminal article was primar-

ily responsible for bringing the general linear model to the attention of social science researchers. The regression approach to ANOVA is accomplished by dummy coding group membership. We will illustrate with two examples that were analyzed in Chapter 2 with traditional ANOVA. The first example had 3 groups, with the following data:

GROUP 1	GROUP 2	GROUP 3
3	4	4
6	7	5
8	9	2
	8	3
		5

We create two dummy variables (DUM1 and DUM2) to identify group membership, and use a 1 on the dummy variable to indicate group membership. The entities in the third group are uniquely identified by 0 and 0 on the two dummy variables, i.e., not in groups 1 or 2. Thus, we have

DEP	DUM1	DUM2	DEP	DUM1	DUM2
3	1	0	4	0	0
6	1	0	5	0	0
8	1	0	2	0	0
4	0	1	3	0	0
7	0	1	5	0	0
9	0	1			
8	0	1			

The second example had four groups, with the following data:

GROUP 1	GROUP2	GROUP3	GROUP4
2	7	4	8
3	9	4	4
5	11	5	7
6		8	7
		3	

In this case we need 3 dummy variables to identify group membership (DUM1, DUM2, and DUM3):

TABLE 6.12
Sample Size Such That the Difference Between the Squared Multiple Correlation and Squared Cross-Validated Correlation
Is Arbitrarily Small With Given Probability

Three Predictors probability										Four Predictors probability						
ρ^2	ϵ	.99	.95	.90	.80	.60	.40	ρ^2	ϵ	.99	.95	.90	.80	.60	.40	
.05	.01	858	554	421	290	158	81	.05	.01	1041	707	559	406	245	144	
	.03	269	166	123	79	39	18		.03	312	201	152	103	54	27	
	.01	825	535	410	285	160	88		.01	1006	691	550	405	253	155	
	.03	271	174	133	91	50	27	.10	.03	326	220	173	125	74	43	
	.05	159	100	75	51	27	14		.05	186	123	95	67	38	22	
.25	.01	693	451	347	243	139	79		.01	853	587	470	348	221	140	
	.03	232	151	117	81	48	27		.03	283	195	156	116	73	46	
	.05	140	91	71	50	29	17	.25	.05	168	117	93	69	43	28	
	.10	70	46	36	25	15	7		.10	84	58	46	34	20	14	
	.20	34	22	17	12	8	6		.20	38	26	20	15	10	7	
.50	.01	464	304	234	165	96	55		.01	573	396	317	236	152	97	
	.03	157	104	80	57	34	21		.03	193	134	108	81	53	35	
	.05	96	64	50	36	22	14	.50	.05	117	82	66	50	33	23	
	.10	50	34	27	20	13	9		.10	60	43	35	27	19	13	
	.20	27	19	15	12	9	7		.20	32	23	19	15	11	9	
.75	.01	235	155	120	85	50	30		.01	290	201	162	121	78	52	
	.03	85	55	43	31	20	13		.03	100	70	57	44	30	21	
	.05	51	35	28	21	14	10	.75	.05	62	44	37	28	20	15	
	.10	28	20	16	13	9	7		.10	34	25	21	17	13	11	
	.20	16	12	10	9	7	6		.20	19	15	13	11	9	7	
.98	.01	23	17	14	11	9	7		.01	29	22	19	15	12	10	
	.03	11	9	8	7	6	6		.03	14	11	10	9	8	7	
	.05	9	7	7	6	6	5	.98	.05	10	9	8	8	7	7	
	.10	7	6	6	6	5	5		.10	8	8	7	7	7	6	
	.20	6	6	5	5	5	5		.20	7	7	7	6	6	6	

		Eight Predictors probability								Fifteen Predictors probability							
ρ^2	ϵ	.99	.95	.90	.80	.60	.40	ρ^2	ϵ	.99	.95	.90	.80	.60	.40		
.05	.01	1640	1226	1031	821	585	418	.05	.01	2523	2007	1760	1486	1161	918		
	.03	447	313	251	187	116	71		.03	640	474	398	316	222	156		
	.01	1616	1220	1036	837	611	450		.01	2519	2029	1794	1532	1220	987		
.10	.03	503	373	311	246	172	121	.10	.03	762	600	524	438	337	263		
	.05	281	202	166	128	85	55		.05	403	309	265	216	159	119		
	.01	1376	1047	893	727	538	404		.01	2163	1754	1557	1339	1079	884		
.25	.03	453	344	292	237	174	129	.25	.03	705	569	504	431	345	280		
	.05	267	202	171	138	101	74		.05	413	331	292	249	198	159		
	.10	128	95	80	63	45	33		.10	191	151	132	111	87	69		
.50	.20	52	37	30	24	17	12	.50	.20	76	58	49	40	30	24		
	.01	927	707	605	494	368	279		.01	1461	1188	1057	911	738	608		
	.03	312	238	204	167	125	96		.03	489	399	355	306	249	205		
.75	.05	188	144	124	103	77	59	.75	.05	295	261	214	185	151	125		
	.10	96	74	64	53	40	31		.10	149	122	109	94	77	64		
	.20	49	38	33	28	22	18		.20	75	62	55	48	40	34		
.98	.01	470	360	308	253	190	150	.98	.01	741	605	539	466	380	315		
	.03	162	125	108	90	69	54		.03	255	210	188	164	135	113		
	.05	100	78	68	57	44	35		.05	158	131	118	103	86	73		
	.10	54	43	38	32	26	22		.10	85	72	65	58	49	43		
	.20	31	25	23	20	17	15		.20	49	42	39	35	31	28		
	.01	47	38	34	29	24	21		.01	75	64	59	53	46	41		
	.03	22	19	18	16	15	14		.03	36	33	31	29	27	25		
	.05	17	16	15	14	13	12		.05	28	26	25	24	23	22		
	.10	14	13	12	12	11	11		.10	23	21	21	20	20	19		
	.20	12	11	11	11	11	10		.20	20	19	19	19	18	18		

*Entries in the body of the table are the sample size such that $P(\rho^2 - p^2c < \epsilon) = \gamma$ where ρ is population multiple correlation, ϵ is some tolerance and γ is the probability.

2	1	0	0
3	1	0	0
5	1	0	0
6	1	0	0
7	0	1	0
9	0	1	0
11	0	1	0
4	0	0	1
4	0	0	1
5	0	0	1
8	0	0	1
3	0	0	1
8	0	0	0
4	0	0	0
7	0	0	0
7	0	0	0

Note, that again the subjects in the last group (4th group here) are identified by 0s on all dummy variables, i.e., not in groups 1, 2, or 3. In general, we need $(k-1)$ dummy variables for k groups.

When the above two data sets were run on SPSS or Windows 7.5 as regression analyses, predicting the dependent variable from group membership (DUM1 and DUM2 were the predictors in the first analysis and DUM1, DUM2, and DUM3 were the predictors in the second analysis), the results were as follows:

Placeholder for unnumbered tables p. 291 of previous edition

(Split table using first two blocks; see instructions on press layout)

Placeholder for unnumbered tables from p. 291 and 292 of previous edition.

Position bottom $\frac{1}{2}$ from element on p. 291, followed by element from p. 292 in this space.

See press layout for details.

Note that the results are *identical* to what was obtained in Chapter 2. The mean square due to regression corresponds to mean square between, while the residual corresponds to mean square error. The mean square due to regression is just variability due to group membership. We will see in the next chapter, on analysis of covariance (which combines ANOVA and regression analysis), that analysis of covariance can be done through regression analysis also.

6.18 SUMMARY OF IMPORTANT POINTS

1. A particularly good situation for multiple regression is where each of the predictors is correlated with y and the predictors have low intercorrelations, for then each of the predictors is accounting for a relatively distinct part of the variance on y .
2. Moderate to high correlations among the predictors (multicollinearity) creates three problems: it (a) severely limits the size of R , (b) makes determining the importance of given predictor difficult, and (c) increases the variance of regression coefficients, making for an unstable prediction equation. One way of combating this problem is to combine into a single measure a set of predictors that are highly correlated.
3. Preselecting a small set of predictors by examining a correlation matrix from a large initial set, or by using one of the stepwise procedures (forward, stepwise, backward) to select a small set, is likely to produce an equation that is sample specific. If one insists on doing this, and I do not recommend it, then the onus is on the investigator to demonstrate that the equation has adequate predictive power beyond the derivation sample.
4. Mallows' C_p was presented as a measure that minimizes the effect of underfitting (important predictors left out of the model) and overfitting (having predictors in the model that make essentially no contribution or are marginal). This will be the case if one chooses models for which $C_p \approx p$.
5. With many data sets, more than one model will provide a good fit to the data. Thus, one deals with selecting a model from a *pool* of candidate models.
6. There are various graphical plots for assessing how well the model fits the assumptions underlying linear regression. One of the most useful graphs the standardized residuals (y axis) versus the predicted values (x axis). If the assumptions are tenable, then one should observe roughly a random scattering. Any *systematic clustering* of the residuals indicates a model violation(s).
7. It is crucial to validate the model(s) by either randomly splitting the sample and cross-validating, or using the PRESS statistic, or by obtaining the Stein estimate of the *average* predictive power of the equation on other samples from the same population. Studies in the literature that have not cross-validated should be

checked with the Stein estimate to assess the generalizability of the prediction equation(s) presented.

8. Results from the Park and Dudycha study indicate that the magnitude of the *population* multiple correlation strongly affects how many subjects will be needed for a reliable prediction equation. If your estimate of the squared population value is .50, then about 15 subjects per predictor are needed. On the other hand, if your estimate of the squared population value is substantially *larger* than .50, then far less than 15 subjects per predictor will be needed. Table 6.8 shows that if $R^2 = .75$, then 10 subjects per predictor will yield a reliable equation. If $R^2 = .85$ (*very strong*) then five subjects per predictor is enough.

9. Influential data points, that is, points that strongly affect the prediction equation, can be identified by seeing which cases have Cook distances >1 . These points need to be examined very carefully. If such a point is due to a recording error, then one would simply correct it and redo the analysis. Or if it is found that the influential point is due to an instrumentation error or that the process that generated the data for that subject was different, then it is legitimate to drop the case from the analysis. If, however, none of these appears to be the case, then one should *not* drop the case, but perhaps report the results of several analyses: one analysis with all the data and an additional analysis(es) with the influential point(s) deleted.

10. It was shown that analysis of variance can be considered as a special case of regression analysis by dummy coding group membership.

EXERCISES

1. Consider this set of data:

x	y
2	3
3	6
4	8
6	4
7	10
8	14
9	8
10	12
11	14
12	12
13	16

- (a) Plot the data. Does there appear to be a linear relationship?
- (b) Run this data on SPSS, obtaining the case analysis.

- (c) Do you see any pattern in the plot of the standardized residuals? What does this suggest?
- (d) Sketch in the regression line, and indicate the raw residuals by vertical lines.

2. Consider the following small set of data:

PREDX	DEP
0	1
1	4
2	6
3	8
4	9
5	10
6	10
7	8
8	7
9	6
10	5

- (a) Plot the points. What type of relationship does this suggest?
- (b) Run this data on SPSS, forcing the predictor in and obtaining the case analysis.
- (c) Do you see any pattern in the plot of the standardized residuals? What does this suggest?

3. Consider the following correlation matrix:

	y	x_1	x_2
y	1.00	.60	.50
x_1	.60	1.00	.80
x_2	.50	.80	1.00

- (a) How much variance on y will x_1 account for if entered first?
- (b) How much variance on y will x_1 account for if entered second?
- (c) What, if anything, do the above results have to do with the multicollinearity problem?
4. A medical school admissions official has two proven predictors (x_1 and x_2) of success in medical school. He has two other predictors under consideration (x_3 and x_4), of which he wishes to choose just one which will add the most (beyond what x_1 and x_2 already predict) to predicting success. Below is the matrix of intercorrelations he has gathered on a sample of 100 medical students:

	x_1	x_2	x_3	x_4
y	.60	.55	.60	.46
x_1		.70	.60	.20
x_2			.80	.30
x_3				.60

- (a) What procedure would he use to determine which predictor has the greater incremental validity? Do *not* go into any numerical details, just indicate the general procedure. Also, what is your educated guess as to which predictor (x_3 or x_4) will probably have the greater incremental validity.
5. Consider the following random sample (in the following table) of about 50% from the Agresti data (in Appendix A in the back of the book) on home sales in Florida. We wish to predict PRICE from the other 4 variables as predictors. The other variables are NEW (whether the home was new or not), NOBATH (number of bathrooms), NOBED (number of bedrooms) and SIZE (size of the house).
- (a) Run stepwise regression analysis on this data. What model is selected?
- (b) Run backward elimination on this data. What model is selected?
6. An investigator has 15 variables on a file. Denote them by $x_1, x_2, x_3, \dots, x_{15}$. Assume there are spaces between all variables, so that free format can be used to read the data. The investigator wishes to predict x_4 . First, however, he obtains the correlation matrix among the predictors and finds that variables 7 and 8 are highly correlated, and decides to combine those as a single predictor. He will also use variables 1, 3, 11, 12, 13, and 14 as predictors. Show the set of control lines for running a stepwise analysis and also obtaining a scatterplot of the residuals vs predicted values of y .
7. A different investigator has 8 variables on a file, with no spaces between the variables, so that fixed format will be needed to read the data. The data looks as follows:

2534674823178659
 3645738234267583
 ETC.

The first two variables are single digit integers, the next three variables are two digit integers, the next two variables are three digit integers and the 8th variable is a two digit integer. The 8th variable is the dependent variable. She wishes to force in variables 1 and 2, and then determine whether variables 3 through 5 (as a block) have any incremental validity. Show the complete SPSS REGRESSION control lines for doing this analysis.

Case Summaries^a

	NEW	NOBATH	NOBED	PRICE	SIZE
1	.00	1.00	3.00	48.50	1.10
2	.00	2.00	3.00	55.00	1.01
3	.00	3.00	3.00	137.00	2.40
4	1.00	3.00	4.00	309.40	3.30
5	.00	1.00	3.00	19.80	1.28
6	.00	1.00	3.00	24.50	.74
7	.00	1.00	2.00	34.80	.78
8	.00	1.00	3.00	32.00	.97
9	.00	1.00	3.00	28.00	.84
10	.00	2.00	2.00	49.90	1.08
11	.00	2.00	3.00	61.50	1.01
12	.00	2.00	3.00	68.90	1.29
13	.00	2.00	3.00	70.50	1.25
14	.00	2.00	3.00	72.90	1.28
15	.00	2.00	3.00	72.00	1.36
16	.00	2.00	3.00	71.00	1.20
17	.00	2.00	3.00	73.00	1.22
18	.00	2.00	2.00	70.00	1.40
19	.00	2.00	2.00	76.00	1.15
20	.00	2.00	3.00	75.50	1.62
21	.00	2.00	3.00	76.00	1.68
22	.00	2.00	3.00	81.80	1.33
23	.00	2.00	3.00	84.50	1.34
24	1.00	2.00	3.00	86.90	1.58
25	.00	2.00	3.00	88.10	2.10
26	.00	2.00	3.00	89.50	1.34
27	.00	2.00	3.00	90.00	1.55
28	1.00	2.00	3.00	95.50	1.54
29	1.00	2.00	4.00	99.90	1.62
30	1.00	2.00	3.00	102.30	1.42
31	1.00	2.00	3.00	110.80	1.56
32	1.00	2.00	3.00	97.90	2.00
33	1.00	2.00	3.00	106.30	1.45
34	.00	2.00	3.00	106.50	1.65
35	1.00	2.00	4.00	109.90	2.06
36	.00	2.00	4.00	110.00	1.76
37	1.00	2.00	4.00	115.00	1.80
38	.00	2.00	3.00	114.90	1.57
39	.00	2.00	4.00	115.00	2.07
40	.00	2.00	4.00	117.90	1.99
41	.00	2.00	3.00	110.00	1.55
42	1.00	2.00	3.00	128.00	1.88
43	1.00	2.00	4.00	139.30	2.05
44	.00	3.00	3.00	142.00	2.12
45	.00	2.00	5.00	148.00	2.40
46	.00	3.00	3.00	150.00	2.04
Total N	46	46	46	46	46

^aLimited to first 100 cases.

8. A regression analysis was run on the Sesame St ($n = 240$) data set, predicting postbody from the following 5 pretest measures: prebody, prelet, preform, prenumb and prerelat. This was run in the syntax editor on SPSS for Windows 12.0. The control lines for doing a stepwise regression, obtaining the 10 largest values for the standardized residuals, the hat elements and Cook's distance, and for obtaining a plot of the standardized residuals versus the predicted y values are given below:

```
TITLE 'MULT REG ON POSTBODY-5 PREDICTORS' .
DATA LIST FREE/ID SITE SEX AGE VIEWCAT SETTING VIEWENC
PREBODY PRELET PREFORM
PRENUMB PRERELAT PRECLASF POSTBODY POSTLET POSTFORM
POSTNUMB POSTREL
POSTCLAS PEABODY.
BEGIN DATA.

DATA LINES.
END DATA.
REGRESSION DESCRIPTIVES = DEFAULT/
VARIABLES = PREBODY TO PRERELAT POSTBODY/
STATISTICS = DEFAULTS TOL SELECTION/
DEPENDENT = POSTBODY/
METHOD = STEPWISE/
RESIDUALS = OUTLIERS (ZRESID, LEVER, COOK) /
SCATTERPLOT (*RES, *PRE) / .
```

The SPSS Windows 7.5 printout follows. Answer the following questions:

- Why did PREBODY enter the prediction equation first?
- Why did PREFORM enter the prediction equation second?
- Write the prediction equation, rounding off to 3 decimals.
- Is multicollinearity present? Explain.
- Compute the Stein estimate and indicate in words exactly what it represents.
- Refer to the standardized residuals. Is the number of these greater than $|2|$ about what you would expect if the model is appropriate? Why, or why not?
- Are there are outliers on the set of predictors?
- Are there any influential data points? Explain.
- From examination of the residual plot, does it appear there may be some model violation(s)? Why, or why not?
- Are the values of VIF (variance inflation factor) for the predictors in the equation reasonable, according to Myers?

(k) Does the value of Mallows prediction criterion for model 2 seem reasonable? What about for model 1?

Placeholder for tinted element p. 299

28.5 pi wide

20 pi deep

Placeholder for element from p. 300 of previous edition

28.5 pi wide

43 pi deep

Placeholder for element from p. 301 of previous edition.

28.5 pi wide

38.5 pi deep

Placeholder for element from p. 302 of previous edition.

28.5 pi wide

34.5 pi deep

Placeholder for element from p. 303 of previous edition.

28.5 pi wide

22 pi deep

9. Show how the partial correlation of .459 is obtained for COUEVAL in MODEL 1 under EXCLUDED VARIABLES for the MORRISON data.
10. Run a stepwise regression analysis for the full AGRESTI data on the CD.
11. Run backward selection on the full AGRESTI data. Do you get the same model?

APPENDIX THE PRESS STATISTIC

As pointed out by several authors, in many instances one does not have enough data to do a random split. One can obtain a good measure of the *external* predictive power by use of the PRESS statistic. In this approach the y value for each subject is set aside and a prediction equation is derived on the remaining data. Thus, n prediction equations are derived and n true prediction errors are found. To be very specific, the prediction error for subject 1 is computed from the equation derived on the remaining $(n - 1)$ data points, the prediction error for subject 2 is computed from the equation derived on the other $(n - 1)$ data points, etc. As Myers (1990) put it, "PRESS is important in that one has information in the form of n validations in which the fitting sample for each is of size $n - 1$ " (p. 171).

The PRESS statistic is especially important when one does not have large sample size, for in this case data splitting is really not practical. For example, if $n = 60$ and we have 6 predictors, randomly splitting the sample involves obtaining a prediction equation on only 30 subjects.

Recall that in deriving the prediction (via the least squares approach), the sum of the squared errors is *minimized*. The PRESS residuals, on the other hand, are true prediction errors, since the y value for each subject was not simultaneously used for fit and model assessment. Let us denote the predicted value for subject i , where that subject was *not* used in developing the prediction equation, by $y_{(-i)}$. Then the PRESS residual for each subject is given by

$$\hat{e}_{(-i)} = y_i - \hat{y}_{(-i)}$$

and the PRESS sum of squared residuals is given by

$$\text{PRESS} = \sum_{(-i)} \hat{e}_{(-i)}^2$$

Therefore, one might prefer the model with the smallest PRESS value. The above PRESS value can be used to calculate an R^2 -like statistic that more accurately reflects the generalizability of the model. It is given by

$$R_{\text{PRESS}}^2 = 1 - (\text{PRESS}) / \sum (y_i - \bar{y})^2$$

Importantly, the SAS REG program does routinely print out PRESS, although it is called PREDICTED RESID SS (PRESS). Given this value, it is a simple matter to calculate the R^2 PRESS statistic, since $s_y^2 = \sum (y_i - \bar{y})^2 / (n - 1)$.

Analysis of Covariance

CONTENTS

- 7.1 Introduction
- 7.2 Purposes of Covariance
- 7.3 Adjustment of Posttest Means
- 7.4 Reduction of Error Variance
- 7.5 Choice of Covariates
- 7.6 Numerical Example
- 7.7 Assumptions in Analysis of Covariance
- 7.8 Use of ANCOVA with Intact Groups
- 7.9 Computer Example for ANCOVA
- 7.10 Alternative Analyses
- 7.11 An Alternative to the Johnson–Neyman Technique
- 7.12 Use of Several Covariates
- 7.13 Computer Example with Two Covariates
- 7.14 Summary

7.1 INTRODUCTION

In Chapter 4 we examined the effect of two or more independent variables (factors) in explaining variation on the dependent variable. We set up an experimental design, and thus this method is called experimental control. In this chapter we consider explaining variation on the dependent variable by measuring the subjects on some other variable(s), called covariates, that are correlated with the dependent variable. Recall that the square of a correlation can be interpreted as “proportion of variance accounted for.” Thus, if we find that I.Q. is correlated with achievement

(dependent variable), say .60, we will be able to attribute 36% of the within group variance on the dependent variable to variability on I.Q. In analysis of covariance (ANCOVA), this part of the variance is removed from the error term, and yields a more powerful test. This method of explaining variation is called statistical control. We now consider an example to illustrate how ANCOVA can be very useful in an experimental study in which the subjects have been randomly assigned to groups.

Example

Suppose an investigator is comparing the effects of two treatments on achievement in science. He assesses achievement through the use of a 50 item multiple choice test. He has 24 students and is able to randomly assign 12 of them to each of the treatments. I.Q. scores are also available for these subjects. The data are as follows:

	Treat. 1		Treat. 2	
	I.Q.	Ach.	I.Q.	Ach.
	100	23	96	19
	113	31	108	26
	98	35	122	31
	110	28	103	22
	124	40	132	36
	135	42	120	38
	118	37	111	31
	93	29	93	25
	120	34	115	29
	127	45	125	41
	115	33	102	27
	104	25	107	21
Means	113.08	33.5	111.17	28.83

The investigator feels no need to use the I.Q. data for analysis purposes since the groups have been “equated” on all variables because of the random assignment. He therefore runs a t test for independent samples on achievement at the .05 level. He finds $t = 1.676$, which is not significant because the critical values are ± 2.074 .

Because of small sample size we have a power problem. The estimated effect size is $\hat{d} = (33.5 - 28.83)/6.83 = .68$ (cf. Section 3.2), which is undoubtedly of practical significance since the groups differ by about two-thirds of a standard deviation. We have not detected it because of the power problem *and* because there is considerable within group variability on achievement. In fact, the pooled within correlation of I.Q. with achievement for the above data is about .80. This means that 64% of the variation in achievement test scores is associated with variation

(individual differences) on I.Q. An analysis of covariance removes that portion from the error term and yields a t value significant at the .05 level ($t = 2.25$). Actually it comes out as an F statistic, so you need to take the square root. Recall that $F = t^2$ for two groups. After reading this chapter, the reader will be able to verify the above t value by running the ANCOVA on SAS or SPSS.

The above example showed that analysis of covariance is very useful in creating a more powerful test in an experimental study. ANCOVA is also used to reduce bias when comparing intact or self-selected groups, such as males and females, Head Start and non-Head Start. A classical use is adjusting posttest means on the dependent variable for any initial differences that may have been present on a pretest. Another typical use is in teaching methods studies that use intact classrooms. If the average I.Q.'s for the classrooms differ by 10 points, then an adjustment of the posttest achievement is done. Although the use of analysis of covariance in this context may seem reasonable, it is quite controversial, which we discuss in detail in Section 7.8.

The first 10 sections of this chapter cover the basics for ANCOVA with one covariate. We discuss the purposes of covariance, the underlying concepts, the assumptions, interpretation of results, the relationship of ANOVA and ANCOVA, and the running of ANCOVA on SAS and SPSS. The last five sections are more advanced, especially the section on the Johnson–Neyman technique, and may be skipped without loss of continuity. Much has been written about analysis of covariance, and the reader should at least be aware of two classic review articles by Cochran (1957) and Elashoff (1969), and a very comprehensive and thorough book on covariance and alternatives by Huitema (1980).

7.2 PURPOSES OF COVARIANCE

Analysis of covariance is related to the following two basic objectives in experimental design:

1. Elimination of systematic bias.
2. Reduction of within group or error variance.

Systematic bias means that the groups differ systematically on some key variable(s) that are related to performance on the dependent variable. If the groups involve treatments, then a significant difference on a posttest at the end of treatments will be confounded (mixed in with) with initial differences on a key variable. It would not be clear whether the treatments were making the difference, or whether initial differences simply transferred to posttest means. A simple example is a teaching methods study with initial differences between groups on I.Q. Suppose two methods of teaching algebra are compared (same teacher for both methods)

with two classrooms in the same school. The following summary data, means for the groups, are available:

	Method 1	Method 2
I.Q.	120.2	105.8
Posttest	73.4	67.5

If the t test for independent samples on the posttest is significant, then it isn't clear whether it was method 1 that made the difference, or the fact that the children in that class were "brighter" to begin with, and thus would be expected to achieve higher scores.

As another example, suppose we are comparing the effect of four stress situations on blood pressure (the dependent variable). It is found that situation 3 is significantly more stressful than the other three situations. However, we note that the blood pressure of the subjects in group 3 under minimal stress is greater than for the subjects in the other groups. Then, it isn't clear that situation 3 is necessarily most stressful. We need to determine whether the blood pressure for group 3 would still be higher if the posttest means for all 4 groups were "adjusted" in some way to account for initial differences in blood pressure. We see later that the posttest means are adjusted in a linear fashion to what they would be if all groups started out equally on the covariate, that is, at the grand mean.

The best way of dealing with systematic bias is to randomly assign subjects to groups. Then we can be confident, within sampling error, that the groups don't differ systematically *on any variables*. Of course, in many studies random assignment is not possible, so we look for ways of at least partially equating groups. One way of partially controlling for initial differences is to match on key variables. Of course, then we can only be sure the groups are equivalent on those matched variables. Analysis of covariance is a *statistical* way of controlling on key variables. Once again, as with matching, ANCOVA can only *reduce* bias, and not eliminate it.

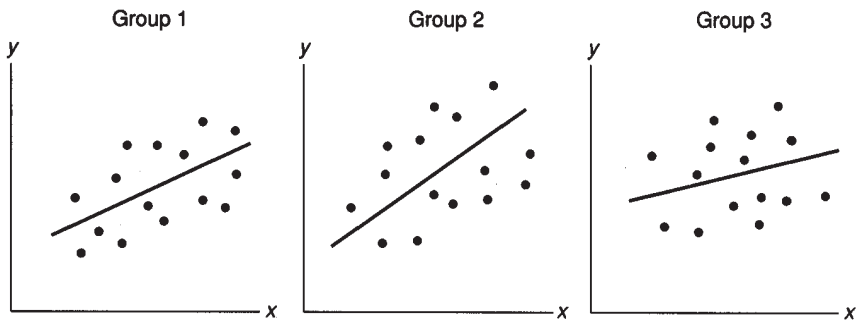
Why is reduction of error variance, the second purpose of analysis of covariance, important? Recall from Chapter 2 on one way ANOVA that the F statistic was $F = MS_b/MS_w$, where MS_w was the estimate of error. If we can make MS_w smaller, then F will be larger and we will obtain a more sensitive or powerful test. And from Chapter 3 on power, remember that power is generally poor in small or medium sample size studies. Thus the use of perhaps 2 or 3 covariates in such studies should definitely be considered. The use of covariates that have relatively low correlations with each other are particularly helpful because each covariate removes a somewhat different part of the error variance from the dependent variable.

Analysis of covariance is a *statistical* way of reducing error variance. There are several other ways of reducing error variance. One way is through sample selec-

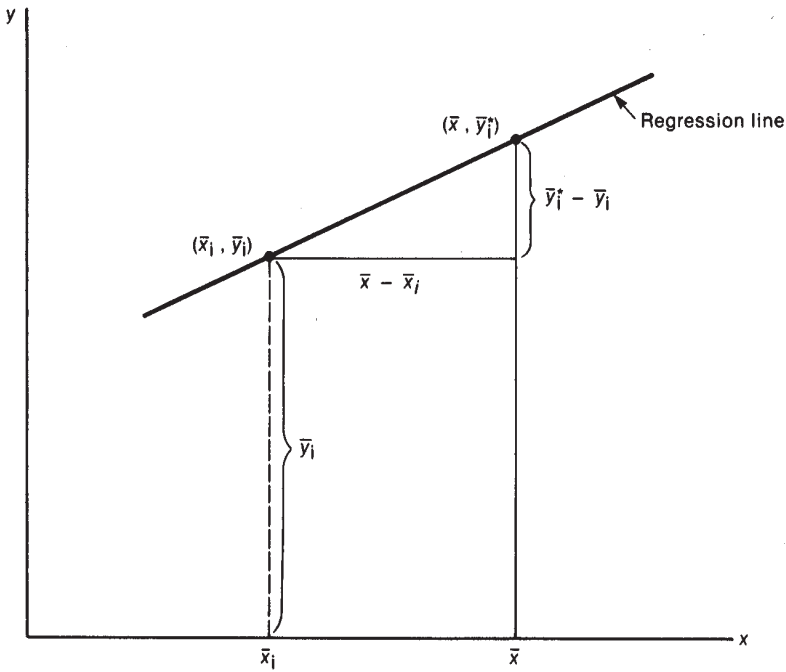
tion; subjects who are more homogeneous vary less on the dependent measure. Another way, discussed in Chapter 4 on factorial designs, was to block on a variable, or consider it as another factor in the design.

7.3 ADJUSTMENT OF POSTTEST MEANS

As mentioned earlier, analysis of covariance adjusts the posttest means to what they would be if all groups started out equally on the covariate; at the grand mean. In this section we derive the general equation for linearly adjusting the posttest means for one covariate. Before we do that, however, it is important to discuss one of the assumptions underlying the analysis of covariance. That assumption for one covariate requires *equal population regression slopes* for all groups. Consider a three group situation, with 15 subjects per group. Suppose that the scatterplots for the 3 groups looked as given below.



Recall from beginning statistics that the x and y scores for each subject determine a point in the plane. Requiring that the slopes be equal is equivalent to saying that the nature of the linear relationship is the *same* for all groups, or that the rate of change in y as a function of x is the same for all groups. For the above scatterplots the slopes are different, with the slope being the largest for group 2 and smallest for group 3. But the issue is whether the *population* slopes are different, and whether the sample slopes differ sufficiently to conclude that the population values are different. With small sample sizes as in the above scatterplots, it is dangerous to rely on visual inspection to determine whether the population values are equal, because of considerable sampling error. Fortunately there is a statistic for this, and later we indicate how to obtain it on SPSS and SAS. In deriving the equation for the adjusted means we are going to assume the slopes are equal. What if the slopes are not equal? Then ANCOVA is *not* appropriate, and we indicate alternatives later on in the chapter.



$$\text{Slope of Straight Line} = b = \frac{\text{change in } y}{\text{change in } x}$$

$$b = \frac{\bar{y}_i^* - \bar{y}_i}{\bar{x} - \bar{x}_i}$$

$$b(\bar{x} - \bar{x}_i) = \bar{y}_i^* - \bar{y}_i$$

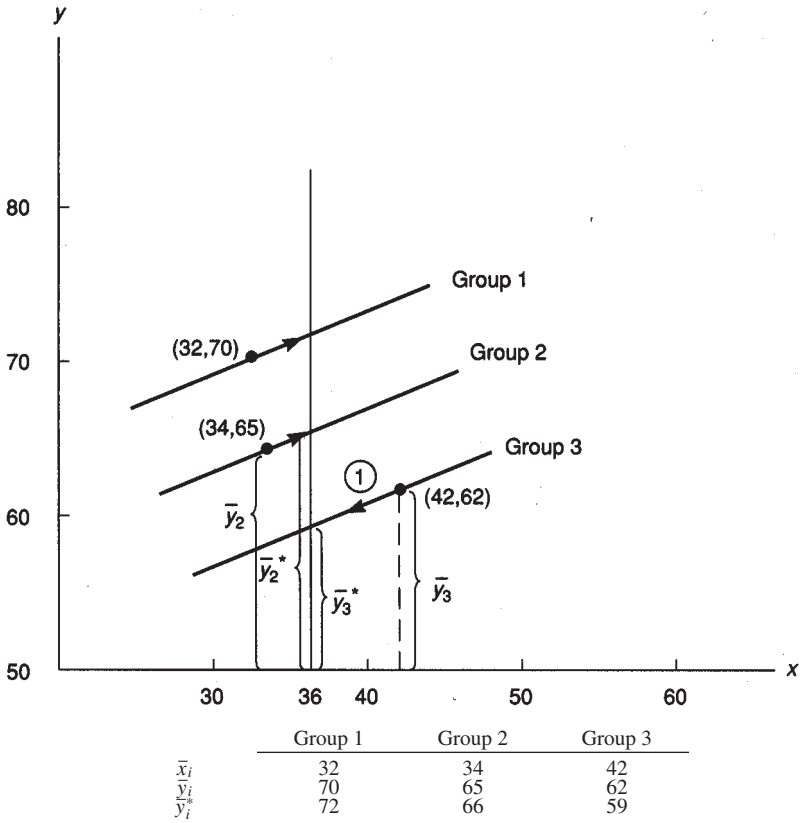
$$\bar{y}_i^* = \bar{y}_i + b(\bar{x} - \bar{x}_i)$$

$$\bar{y}_i^* = \bar{y}_i - b(\bar{x}_i - \bar{x})$$

FIGURE 7.1 Deriving the General Equation for Adjusted Means in Covariance

The details of obtaining the adjusted mean for the i th group (i.e., any group) are given in Figure 7.1. The general equation follows from the definition for the slope of a straight line and some basic algebra.

In Figure 7.2 we show the adjusted means geometrically for a hypothetical 3 group data set. A positive correlation is assumed between the covariate and the dependent variable, so that a higher mean on x implies a higher mean on y . Note that since group 1 scored below the grand mean on the covariate, its mean is adjusted upward. On the other hand, since the mean for group 3 on the covariate is *above* the



A common slope = .5 is being assumed here.

$$\bar{y}_1^* = 70 - .5(32 - 36), \bar{y}_2^* = 65 - .5(34 - 36), \bar{y}_3^* = 62 - .5(42 - 36)$$

① The arrows on the regression lines indicate that the means are adjusted linearly upward or downward to what they would be if the groups had started out at the grand mean on the covariate.

FIGURE 7.2 Means and Adjusted Means for Hypothetical Three Group Data Set

grand mean, covariance estimates that it would have scored lower on y if its mean on the covariate was lower (at grand mean), and therefore the mean for group 3 is adjusted downward.

7.4 REDUCTION OF ERROR VARIANCE

It is relatively simple to derive the approximate error term for covariance. Denote the correlation between the covariate (x) and the dependent variable (y) by r_{xy} . The square of a correlation can be interpreted as “proportion of variance accounted for.” The within group variance for ANOVA is MS_w . Thus, the part of the within group variance on y that is accounted for by the covariate is $r_{xy}^2 MS_w$. The within variability left, after the portion due to the covariate is removed, is

$$MS_w - MS_w r_{xy}^2 = MS_w (1 - r_{xy}^2) \quad (1)$$

and this becomes our new error term for the analysis of covariance, which we denote by MS_w^* . Technically, there is an additional part to the adjusted error term:

$$MS_w^* = MS_w (1 - r_{xy}^2) [1 - 1/(f_e - 2)]$$

where f_e is the error degrees of freedom. However, the effect of this additional factor is slight as long as $N > 50$.

To show how much of a difference a covariate can make in increasing the sensitivity of an experiment, we consider a hypothetical study. An investigator runs a one-way ANOVA (3 groups and 20 subjects per group), and obtains $F = 200/100 = 2$, which is not significant, because the critical value at .05 is 3.18. He pretested the subjects, but didn't use the pretest as a covariate (even though the correlation between covariate and posttest was .71) because the groups didn't differ significantly on the pretest. This is a common mistake made by some researchers who are unaware of the other purpose of covariance, that of reducing error variance. The analysis is redone by another investigator using ANCOVA. Using the equation we just derived she finds

$$MS_w^* \approx 100[1 - (.71)^2] = 50$$

Thus, the error term for the ANCOVA is only half as large as the error term for ANOVA. It is also necessary to obtain a new MS_b^* for ANCOVA, call it MS_b^* . In Section 7.6 we show how to calculate MS_b^* . Let us assume here that the investigator obtains the following F ratio for the covariance analysis:

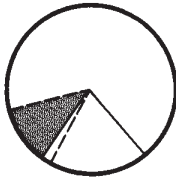
$$F^* = 190/50 = 3.8$$

This is significant at the .05 level. Therefore, the use of covariance can make the difference between finding and not finding significance. Finally, we wish to note that MS_b^* can be smaller or larger than MS_b although in a randomized study the expected values of the two are equal.

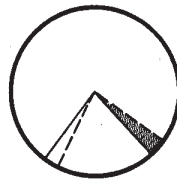
7.5 CHOICE OF COVARIATES

In general, any variables that theoretically should correlate with the dependent variable, or variables that have been shown to correlate on similar types of subjects, should be considered as possible covariates. The ideal is to choose as covariates variables that of course are significantly correlated with the dependent variable *and* have low correlations among themselves. If two covariates are highly correlated (say .80), then they are removing much of the *same* error variance from y ; x_2 will not have much incremental validity. On the other hand, if two covariates (x_1 and x_2) have a low correlation (say .20), then they are removing relatively distinct pieces of the error variance from y , and we will obtain a much greater total error reduction. This is illustrated graphically below using Venn diagrams, where the circle represents error variance on y .

x_1 and x_2 Low correl.



x_1 and x_2 High correl.



Solid lines—part of variance on y that x_1 accounts for.

Dashed lines—part of variance on y that x_2 accounts for.

The shaded portion in each case represents the incremental validity of x_2 , that is, the part of error variance on y it removes that x_1 did not.

Huitema (1980, p. 161) has recommended limiting the number of covariates to the extent that the ratio

$$\frac{[C + (J - 1)]}{N} < .10$$

where C is the number of covariates, J is the number of groups, and N is total sample size. Thus, if we had a four group problem with a total of 80 subjects, then $(C + 3)/80 < .10$ or $C < 5$. Less than 5 covariate should be used. If the above ratio is $> .10$, then the adjusted means are likely to be unstable.

7.6 NUMERICAL EXAMPLE

We now consider an example to illustrate how to calculate an ANCOVA and to make clear what the null hypothesis is that is being tested. We use the following 3 group data set from Myers (1979, p. 417), where x indicates the covariate:

Group 1		Group 2		Group 3	
x	y	x	y	x	y
12	26	11	32	6	23
10	22	12	31	13	35
7	20	6	20	15	44
14	34	18	41	15	41
12	28	10	29	7	28
11	26	11	31	9	30

Recall that in the one way ANOVA the null hypothesis was $H_0: \mu_1 = \mu_2 = \cdots = \mu_k$ (population means are equal). But in analysis of covariance we are adjusting the means (Section 7.3), so that the null hypothesis becomes $H_0: \mu_1^* = \mu_2^* = \cdots = \mu_k^*$, that is, the *adjusted* population means are equal. In the above example, the specific null hypothesis is $H_0: \mu_1^* = \mu_2^* = \mu_3^*$. In ANCOVA we adjust sums of squares corresponding to the sums of squares total, within and between from ANOVA. We denote these adjusted sums of squares by SS_t^* , SS_w^* , and SS_b^* respectively. SS_b^* is obtained by subtracting SS_w^* from SS_t^* .

An ANOVA on the above Myers data, as the reader should check, yields a within cells sum of squares of 666.83 and a group sum of squares of 172.11. We will need these results in obtaining the ANCOVA. Recall that SS_t from ANOVA measures variability of the subjects scores about the grand mean;

$$SS_t = \sum (x_{ij} - \bar{x})^2$$

Let r_{xy} denote the correlation between the covariate and the dependent variable for all the scores, disregarding group membership. Remember that r_{xy}^2 can be interpreted as proportion of variance accounted for. Thus, $r_{xy}^2 SS_t$ represents the amount of variability on y that is accounted for by its relationship with the covariate. Therefore, the remaining variability on y , or the adjusted total sum of squares, is given by

$$SS_t^* = (1 - r_{xy}^2) SS_t \quad (2)$$

Now consider the pooled within correlation for x and y , that is, where group membership is taken into account. Although not strictly true, this correlation can be thought of as the average (or weighted average for unequal group sizes) of the correlations within the groups. Denote this correlation by $r_{xy(w)}$. Then the amount of within group variability on y accounted for by the covariate is given by $r_{xy(w)}^2 SS_w$. Therefore, the remaining within variability on y , or the adjusted within sum of squares, is given by

$$SS_w^* = (1 - r_{xy(w)}^2) SS_w \quad (3)$$

Finally, the adjusted between sum of squares is obtained as the difference between the adjusted total and adjusted within:

$$SS_b^* = SS_t^* - SS_w^* \quad (4)$$

The F ratio for analysis of covariance is then given by

$$F^* = (SS_b^* / (k - 1)) / SS_w^* / (N - K - C) = MS_b^* / MS_w^* \quad (5)$$

where C is the number of covariates. Note that one degree of freedom for error is lost for each covariate used.

This method of computing the ANCOVA is conceptually fairly simple, and importantly shows its direct linkage with the results from an ANOVA on the same data. The SAS GLM control lines for running the ANCOVA are presented in Table 7.1, along with selected printout. The total correlation is .85286 and the within group correlations are gp 1: .9316, gp 2: .9799, and gp 3: .9708. Using these results, the F ratio for the ANCOVA is easily obtained. First, from Equation 2 we have that

$$SS_t^* (1 - (.85286)^2) 838.94 = 228.72$$

TABLE 7.1
SAS GLM Control Lines and Selected Printout for ANCOVA on Myers
Data and SPSS Windows 12.0 Interactive Plots and Regression Lines

Placeholder for T0701 from p. 315 of previous edition

TABLE 7.1
(Continued)

Place holder for T0701-b from p. 316 of previous edition.

Now, using the average of the within correlations as a rough estimate of the pooled within correlation, we find that $r = (.9316 + .9799 + .9708)/3 = .9608$ (the actual pooled correlation is .965). Now, using Equation 3 we find the adjusted within sum of squares:

$$SS_w^* - (1 - (.965)^2)(666.83) = 45.86$$

Therefore, the adjusted between sum of squares is:

$$SS_b^* - 228.72 - 45.86 = 182.86$$

and the F ratio for the analysis of covariance is

$$F^* = (182.86/2)/45.86/(18-3-1) = 27.87$$

ANCOVA as a Special Case of Multiple Regression

Since analysis of covariance involves both analysis of variance and regression analysis, we can do an ANCOVA using multiple regression. Recall that in the last chapter on regression analysis we showed that ANOVA was a special case of regression analysis. We dummy coded group membership and used these dummy variables to predict the dependent variable.

We will illustrate how an ANCOVA can be done using multiple regression with the Myers data. First, we shall check the homogeneity of regression slopes assumption. For a one way design, as Myers and Well (1991, p. 567) point out, “Performing an ANCOVA on a design that has a single factor A can now be seen as determining whether A has effects over and above those of the covariate x . “Thus, we *force* the covariate in and then determine whether group membership has an effect above and beyond the covariate. Since we have 3 groups here, we will need two dummy variables to code group membership (we denote them by DUM1, DUM2). Recall that a violation of the slopes assumption meant there was a group by covariate interaction. Thus, we set up an interaction effect and test it for significance. We create the group by covariate interaction effects by multiplying (we denote them by COVDUM1 and COVDUM2) and then test these for significance. The complete control lines for testing homogeneity of slopes and doing the ANCOVA are presented in Table 7.2.

Selected printout from SPSS for Windows 12.0 is presented in Table 7.3. Note that the assumption of equal regression slopes is tenable ($F = .354$), and that the ANCOVA is significant ($F = 27.886$).

7.7 ASSUMPTIONS IN ANALYSIS OF COVARIANCE

Analysis of covariance rests on the same assumptions as the analysis of variance *plus* three additional assumptions regarding the regression part of the covariance analysis. ANCOVA also assumes

1. A linear relationship between the dependent variable and the covariate(s).

TABLE 7.2
Syntax Command File for Homogeneity of Slopes Test and ANCOVA on Myers Data Using SPSS for Windows 12.0

TITLE 'MULT. REG ON MYERS DATA-ANCOVA'.																																			
DATA LIST FREE/COVAR DEP DUM1 DUM2 COVDUM1 COVDUM2.																																			
BEGIN DATA.																																			
12	26	1	0	12	0	10	22	1	0	10	0	7	20	1	0	7	0	12	28	1	0	12	0	11	26	1	0	11	0						
11	32	0	1	0	11	12	31	0	1	0	12	6	20	0	1	0	6	18	41	0	1	0	18	10	29	0	1	0	10	11	31	0	1	0	11
6	23	0	0	0	0	13	35	0	0	0	0	15	44	0	0	0	0	15	41	0	0	0	0	7	28	0	0	0	0	9	30	0	0	0	
END DATA.																																			
LIST.																																			
REGRESSION DESCRIPTIVES = DEFAULT/																																			
VARIABLES = COVAR TO COVDUM2/																																			
DEPENDENT = DEP/																																			
① ENTER COVAR DUM1 DUM2/TEST (COVDUM1 COVDUM2) /.																																			
REGRESSION DESCRIPTIVES = DEFAULT/																																			
VARIABLES = COVAR TO DUM2/																																			
DEPENDENT = DEP/																																			
② ENTER COVAR/TEST(DUM1 DUM2) /.																																			

① This is the statement which yields the homogeneity of slopes test.
② This is testing the main hypothesis in ANCOVA, whether the adjusted population means are equal.

COVAR	DEP	DUM1	DUM2	COVDUM1	COVDUM2
12.00	26.00	1.00	.00	12.00	.00
10.00	22.00	1.00	.00	10.00	.00
7.00	20.00	1.00	.00	7.00	.00
14.00	34.00	1.00	.00	14.00	.00
12.00	28.00	1.00	.00	12.00	.00
11.00	26.00	1.00	.00	11.00	.00
11.00	32.00	.00	1.00	.00	11.00
12.00	31.00	.00	1.00	.00	12.00
6.00	20.00	.00	1.00	.00	6.00
18.00	41.00	.00	1.00	.00	18.00
10.00	29.00	.00	1.00	.00	10.00
11.00	31.00	.00	1.00	.00	11.00
6.00	23.00	.00	.00	.00	.00
13.00	35.00	.00	.00	.00	.00
15.00	44.00	.00	.00	.00	.00
15.00	41.00	.00	.00	.00	.00
7.00	28.00	.00	.00	.00	.00
9.00	30.00	.00	.00	.00	.00

2. Homogeneity of the regression slopes (for one covariate); parallelism of the regression planes for two covariates and for more than 2 covariates homogeneity of the regression hyperplanes.
3. The covariate is measured without error.

Since covariance rests on the same assumptions as ANOVA, any violations that are serious in ANOVA (like dependent observations) are also serious in ANCOVA. Violation of *all 3* of the above regression assumptions can also be serious. For example, if the relationship between the covariate and the dependent variable is curvilinear, then the adjustment of the means will be improper.

There is always measurement error for the variables that are typically used as covariates in social science research. In randomized designs this reduces the power of the ANCOVA, but treatment effects are not biased. For non-randomized designs the treatment effects can be seriously biased.

A violation of the homogeneity of regression slopes can also yield quite misleading results. To illustrate this, we present in Figure 7.3 the situation where the assumption is met and two situations where the slopes are unequal. Notice that with equal slopes the estimated superiority of group 1 at the grand mean is a totally accurate estimate of group 1's superiority for *all* levels of the covariate, since the lines are parallel. For Case 1 of unequal slopes there is a *covariate by treatment interaction*. That is, how much better group 1 is depends on which value of the covariate we specify. This is analogous to the concept of interaction in a factorial design. For Case 2 of heterogeneous slopes the use of covariance would be totally misleading. Covariance estimates no difference between the groups, while for $x = c$, group 2 is quite superior, and for $x = d$, group 1 is quite superior. Later in the chapter we show how to test the assumption of equal slopes on SPSS and on SAS.

Therefore, in examining printout from the statistical packages it is important to *first* make two checks to determine whether analysis of covariance is appropriate:

1. Check to see whether there is a linear relationship between the dependent variable and the covariate.
2. Check to determine whether the homogeneity of the regression slopes is tenable.

If the above assumptions are met, then there is not any debate about the appropriateness of ANCOVA in randomized studies in which the subjects have been randomly assigned to groups. For intact groups, there is a debate, and we discuss that in the next section.

If either of the above assumptions is not satisfied, then covariance is not appropriate. In particular, if (2) is not met, then one should consider using the

TABLE 7.3
Selected Printout From SPSS for Windows 12.0 for ANCOVA on Myers
Data Using Multiple Regression

- ① This is the test for a significant regression on y on the covariate.
- ② This is the test for homogeneity of the regression slopes.
- ③ This is the main test in covariance; whether the adjusted population means are equal.

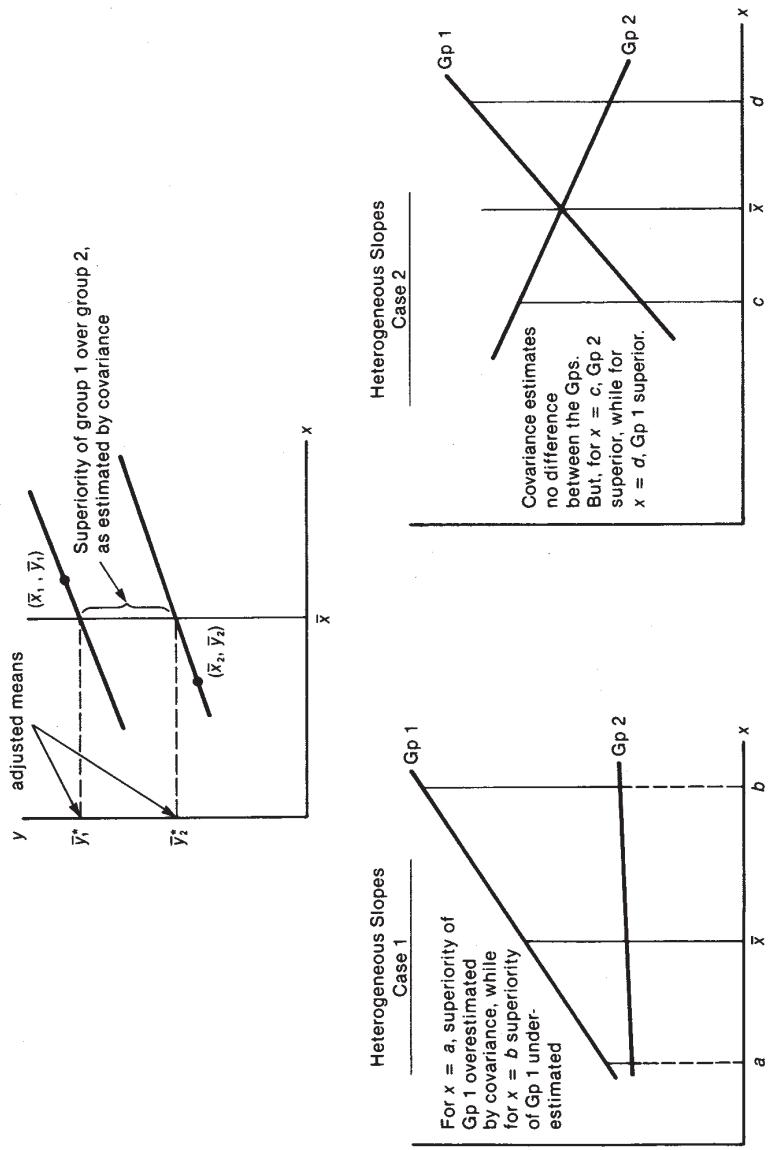


FIGURE 7.3 Effect of Heterogeneous Slopes on Interpretation in ANCOVA

Johnson–Neyman (1936) technique. For extended discussion on the Johnson–Neyman technique see Rogosa (1977, 1980).

7.8 USE OF ANCOVA WITH INTACT GROUPS

It should be noted that some researchers (Anderson, 1963; Lord, 1969) have argued strongly against using analysis of covariance with intact groups. Although we do not take this position, it is important that the reader be aware of the several limitations and/or possible dangers when using ANCOVA with intact groups. First, even the use of several covariates will *not* equate intact groups, and one should never be deluded into thinking it can. The groups may still differ on some unknown important variable(s). Also, note that equating groups on one variable may result in accentuating their differences on other variables.

Second, recall that ANCOVA adjusts the posttest means to what they would be if all the groups had started out equal on the covariate(s). You then need to consider whether groups that are equal on the covariate would ever exist in the real world. Elashoff (1969) gives the following example. Teaching methods *A* and *B* are being compared. The class using *A* is composed of high ability students, whereas the class using *B* is composed of low ability students. A covariance analysis can be done on the posttest achievement scores holding ability constant, as if *A* and *B* had been used on classes of equal and average ability. But, as Elashoff notes, “It may make no sense to think about comparing methods *A* and *B* for students of average ability, perhaps each has been designed specifically for the ability level it was used with, or neither method will, in the future, be used for students of average ability” (p. 387).

Third, the assumptions of linearity and homogeneity of regression slopes need to be satisfied for ANCOVA to be appropriate.

A fourth issue that can confound the interpretation of results is differential growth of subjects in intact or self selected groups on some dependent variable. If the natural growth is much greater in one group (treatment) than for the control group and covariance finds a significance difference, after adjusting for any pretest differences, then it isn’t clear whether the difference is due to treatment, differential growth, or part of each. Bryk and Weisberg (1977) discuss this issue in detail and propose an alternative approach for such growth models.

A fifth problem is that of measurement error. Of course this same problem is present in randomized studies. But there the effect is merely to attenuate power. In non-randomized studies measurement error can seriously bias the treatment effect. Reichardt (1979), in an extended discussion on measurement error in ANCOVA, states,

Measurement error in the pretest can therefore produce spurious treatment effects when none exist. But it can also result in a finding of no intercept difference when a true treatment effect exists, or it can produce an estimate of the treatment effect which is in the opposite direction of the true effect. (p. 164)

It is no wonder then that Pedhazur (1982, p. 524), in discussing the effect of measurement error when comparing intact groups, says,

The purpose of the discussion here was only to alert you to the problem in the hope that you will reach two obvious conclusions: (1) that efforts should be directed to construct measures of the covariates that have very high reliabilities and (2) that ignoring the problem, as is unfortunately done in most applications of ANCOVA, will not make it disappear. (p. 524)

Porter (1967) has developed a procedure to correct ANCOVA for measurement error, and an example illustrating that procedure is given in Huitema (1980, pp. 315–316). This is beyond the scope of the present text.

Given all of the above problems, the reader may well wonder whether we should abandon the use of covariance when comparing intact groups. But other statistical methods for analyzing this kind of data (such as matched samples, gain score ANOVA) suffer from many of the same problems, such as seriously biased treatment effects. The fact is that inferring cause-effect from intact groups is treacherous, regardless of the type of statistical analysis. Therefore, the task is to do the best we can and exercise considerable caution, or as Pedhazur (1982) put it: “But the conduct of such research, indeed all scientific research, requires sound theoretical thinking, constant vigilance, and a thorough understanding of the potential and limitations of the methods being used” (p. 525).

7.9 COMPUTER EXAMPLE FOR ANCOVA

To illustrate how to run an ANCOVA, while at the same time checking the critical assumptions of linearity and homogeneity of slopes, we consider part of a Sesame Street data set from Glasnapp and Poggio (1985), who present data on many variables, including 12 background variables and 8 achievement variables for 240 subjects. Sesame Street was developed as a television series aimed mainly at teaching preschool skills to 3- to 5-year-old children. Data was collected at 5 different sites on many achievement variables both before (pretest) and after (posttest) viewing of the series. We consider here only the achievement variable of knowledge of numbers. The maximum possible score is 54 and the content of the items included recognizing numbers, naming numbers, counting, addition, and subtraction. We use ANCOVA to determine whether the posttest knowledge of numbers for the

TABLE 7.4
SPSS MANOVA Control Lines for Analysis of Covariance on Sesame
Street Data

```

TITLE 'ANALYSIS OF COVARIANCE ON SESAME DATA'.
DATA LIST FREE/SITE PRENUMB POSTNUMB.
BEGIN DATA.

      DATA (ON CD)
END DATA.
MANOVA PRENUMB POSTNUMB BY SITE(1,3) /
① ANALYSIS POSTNUMB WITH PRENUMB /
② PRINT = PMEANS /
  DESIGN /
③ ANALYSIS = POSTNUMB /
  DESIGN = PRENUMB,SITE,PRENUMB BY SITE /
④ ANALYSIS = PRENUMB/.

```

- ① The covariate(s) follow the keyword WITH.
- ② This PRINT subcommand is needed to obtain the adjusted means, which is what we are testing for significance.
- ③ This ANALYSIS subcommand and the following DESIGN subcommand are needed to test the homogeneity of the regression slopes assumption.
- ④ This ANALYSIS subcommand is used to test whether the sites differed significantly on the pretest.

children at the first 3 sites differed after adjustments are made for any pretest differences.

In Table 7.4 we give the complete control lines for running the ANCOVA on SPSS MANOVA. Table 7.5 gives selected annotated output from that run. We indicate *which* of the F tests are checking the assumptions of linearity and homogeneity of slopes, and which F addresses the main question in covariance (whether the adjusted population means are equal).

7.10 ALTERNATIVE ANALYSES

When comparing two or more groups with pretest and posttest data, the following other modes of analysis have been used by many researchers:

1. An ANOVA is done on the difference or gain scores (posttest—pretest).
2. A two way repeated measures (this is covered in Chapter 5) ANOVA is done. This is also called a one between (the grouping variable) and one within (pretest-posttest part) factor ANOVA.

TABLE 7.5
Selected Printout from SPSS MANOVA for ANCOVA on Sesame Street
Data

Placeholder for T0705 from p. 236 of the previous edition.

① This indicates there is a significant correlation between the dependent variable and the covariate(PRENUMB), or equivalently a significant regression of POSTNUMB on PRENUMB.

② This test indicates that homogeneity of regression slopes is tenable at the .05 level, since the p value is .607.

③ This F is testing the main result in ANCOVA; whether the adjusted population means are equal. This is rejected at the .05 level, indicating SITE differences.

④ These are the adjusted means. Since the estimated common regression slope is .686 (given on the printout but not presented here), the adjusted mean for SITE 1 is

$$\bar{y}_1^* = 30.083 - .686(22.4 - 21.67) = 29.58$$

⑤ This test indicates the subjects at the 3 sites differ significantly on the pretest, i.e., on PRENUMB.

Huck and McLean (1975) and Jennings (1988) have compared the above two modes of analysis along with the use of ANCOVA for the pretest-posttest control group design, and conclude that ANCOVA is the preferred method of analysis. Several comments from the Huck and McLean article are worth mentioning. First, they note that with the repeated measures approach it is the *interaction F* that is indicating whether the treatments had a differential effect, and not the treatment main effect. We consider two patterns of means below to illustrate.

	<i>Situation 1</i>			<i>Situation 2</i>	
	Pretest	Posttest		Pretest	Posttest
Treat.	70	80	Treat	65	80
Control	60	70	Control	60	68

In situation 1 the treatment main effect would probably be significant, because there is a difference of 10 in the row means. However, the difference of 10 on the posttest just transferred from an initial difference of 10 on the pretest. There is not a differential change in the treatment and control groups here. On the other hand, in situation 2 even though the treatment group scored higher on the pretest, it increased 15 points from pre to post while the control group increased just 8 points. That is, there was a *differential* change in performance in the two groups. But, recall from Chapter 4 that one way of thinking of an interaction effect is as a “difference in the differences.” This is exactly what we have in situation 2, hence a significant interaction effect.

Second, Huck and McLean (1975) note that the interaction *F* from the repeated measures ANOVA is *identical* to the *F* ratio one would obtain from an ANOVA on the gain (difference) scores. Finally, whenever the regression coefficient is not equal to 1 (generally the case), the error term for ANCOVA will be smaller than for the gain score analysis and hence the ANCOVA will be a more sensitive or powerful analysis.

Although not discussed in the Huck and McLean paper, we would like to add a measurement caution against the use of gain scores. It is a fairly well known measurement fact that the reliability of gain (difference) scores is generally not good. To be more specific, *as the correlation between the pretest and posttest scores approaches the reliability of the test, the reliability of the difference scores goes to 0*. The following table from Thorndike and Hagen (1977) quantifies things:

Correlation between tests	Average reliability of two tests					
	.50	.60	.70	.80	.90	.95
.00	.50	.60	.70	.80	.90	.95
.40	.17	.33	.50	.67	.83	.92
.50	.00	.20	.40	.60	.80	.90
.60		.00	.25	.50	.75	.88
.70			.00	.33	.67	.83
.80				.00	.50	.75
.90					.00	.50
.95						.00

If our dependent variable is some noncognitive measure, or a variable derived from a nonstandardized test (which could well be of questionable reliability), then a reliability of about .60 or so is a definite possibility. In this case, if the correlation between pretest and posttest is .50 (a realistic possibility), the reliability of the difference scores is only .20! On the other hand, the above table also shows that if our measure is quite reliable (say .90), then the difference scores will be reliable for moderate pre-post correlations. For example, for reliability = .90 and pre-post correlation = .50, the reliability of the differences scores is .80.

7.11 AN ALTERNATIVE TO THE JOHNSON–NEYMAN
 TECHNIQUE

We consider hypothetical data from Huitema (1980, p. 272). The effects of two types of therapy are being compared on an aggressiveness score. The covariate (x) are scores on a sociability scale. Since the Johnson–Neyman technique is still (18 years after the first edition of this text) not available on SAS or SPSS, we consider an alternative analysis that does shed some light. Recall that a violation of the homogeneity of regression slopes assumption meant there was a covariate by group interaction. Thus, one way of investigating the nature of this interaction would be to set up a factorial design, with groups being one of the factors and two or more levels for the covariate (other factor), and run a regular two way ANOVA. This procedure is not as desirable as the Johnson–Neyman technique for two reasons: (1) the Johnson–Neyman technique is more powerful, and (2) the Johnson–Neyman technique enables us to determine where the group differences are for *all* levels of the covariate, whereas the factorial approach can only check for differences for the levels of the covariate included in the design. Nevertheless, at least most researchers can easily do a factorial design, and this does yield useful information.

For the Huitema data, although there is a strong linear relationship in each group, the assumption of equality of slopes is not tenable (Figure 7.4 shows why). Therefore, covariance is not appropriate, and we split the subjects into three levels

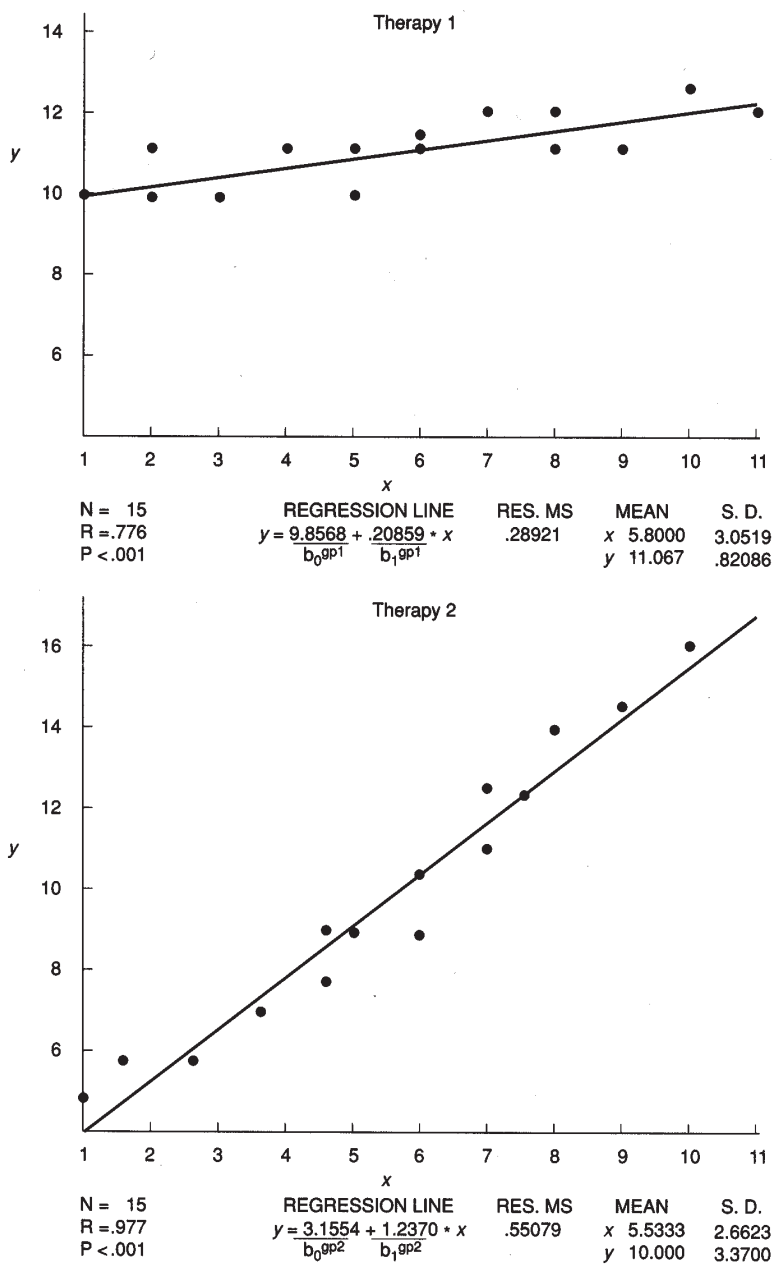


FIGURE 7.4 Scatterplots and Summary Statistics for Each Therapy Group

TABLE 7.6
Results From SPSS for Windows 12.0
for 2×3 Factorial Design on Huitema Data

Placeholder for T0706 from p. 331 of previous edition

for sociability: low (1–4), medium (4.5–7.5) and high (8–11), and set up the following 2×3 ANOVA on aggressiveness:

	SOCIABILITY		
	LOW	MEDIUM	HIGH
THERAPY 1			
THERAPY 2			

Results from the resulting run on SPSS for Windows 12.0 are presented in Table 7.6. They show, as expected, that there is a significant sociability by therapy interaction ($F = 19.735$). The nature of this interaction can be gauged by examining the means for the SOCIAL*THERAPY table. These show that for low sociability therapy group 1 is more aggressive, whereas for high sociability therapy group 2 is more aggressive. The results from the Johnson–Neyman analysis for this data, presented in the first edition of this text (p. 179), show that more precisely there is *no* significant difference in aggressiveness for sociability scores between 6.04 and 7.06.

7.12 USE OF SEVERAL COVARIATES

What is the rationale for using several covariates? First, the use of several covariates will result in greater error reduction than can be obtained with just one covariate. The error reduction will be substantially greater if there are low intercorrelations among the covariates. In this case each of the covariates will be removing a somewhat different part of the error variance from the dependent variable. Also, with several covariates we can make a better adjustment for initial differences among groups.

Recall that with one covariate simple linear regression was involved. With several covariates (predictors), multiple regression is needed. In multiple regression the linear combination of the predictors that is *maximally correlated* with the dependent variable is found. The multiple correlation (R) is a maximized Pearson correlation between the observed scores on y and their predicted scores, $R = r_{\hat{y}y}$. Although R is more complex it is a correlation and hence R^2 can be interpreted as “proportion of variance accounted for.” Also, we will have regression coefficients for each of the covariates (predictors). Below we present a table comparing the single and multiple covariate cases:

TABLE 7.7
SPSS MANOVA Control Lines for ANCOVA on Sesame Street Data
With Two Covariates

TITLE 'SESAME ST. DATA-2 COVARIATES'.		
DATA LIST FREE/ SITE PRENUMB PRERELAT POSTNUMB.		
BEGIN DATA.		
DATA LINES		
END DATA.		
MANOVA PRENUMB PRERELAT POSTNUMB BY SITE(1,3) /		
ANALYSIS POSTNUMB WITH PRENUMB PRERELAT/		
PRINT = PMEANS/		
DESIGN/		
ANALYSIS = POSTNUMB/		
DESIGN = PRENUMB+PRERELAT,SITE,PRENUMB BY SITE+		
PRERELAT BY SITE/		
ANALYSIS = PRENUMB PRERELAT/.		

	One Covariate	Multiple Covariates
Error Reduction	primarily determined by simple correlation r_{yx}^2 – within variance on y accounted for by x	determined by the multiple correlation R^2 – within variance on y accounted for by the set of covariates
Adjustment of Means	$y_i^* = y_i - b(\bar{x}_i - \bar{x})$, b is assumed common slope	$y_{j^*} = y_j - b_1(\bar{x}_{1j} - \bar{x}_1) - b_2(\bar{x}_{2j} - \bar{x}_2) - \cdots - b_k(\bar{x}_{kj} - \bar{x}_k)$

where the b_i are the regression coefficients, \bar{x}_{1j} is the mean for covariate 1 in group j , \bar{x}_{2j} is the mean for covariate 2 in group j , etc., and the \bar{x}_i are the grand means for the covariates.

7.13 COMPUTER EXAMPLE WITH TWO COVARIATES

To illustrate running an ANCOVA with more than one covariate, we reconsider the Sesame Street data set used in Section 7.9. Again we shall be interested in site differences on POSTNUMB, but now we use *two* covariates: PRENUMB and PRERELAT (pretest on knowledge of relational terms—amount, size, and position relationship—maximum score of 17). Before we give the control lines for running the analysis, we need to discuss in more detail how to set up the lines for testing the homogeneity assumption. For one covariate this is equality of regression slopes. For two covariates it is parallelism of the regression planes, and for more than two covariates it involves equality of regression hyperplanes.

TABLE 7.8
Printout from SPSS MANOVA for Sesame Data with Two Covariates

Placeholder for T 0708 from p. 334 of previous edition.

-
- ① This test indicates significant SITE differences at .05 level.
 - ② These are the regression coefficients
 - ③ These are the adjusted means, which would be obtained as follows:

$$\begin{aligned}
 \bar{y}_3^* &= \bar{y}_3 - b_1(\bar{x}_{13} - \bar{x}_1) - b_2(\bar{x}_{23} - \bar{x}_2) \\
 &= 25.437 - .564(16.563 - 21.670) - .622(8.563) - 10.14) \\
 &= 29.30
 \end{aligned}$$

- ④ This test indicates parallelism of the regression planes is tenable at the .05 level.

It is important to recall that a violation of the assumption means there is a covariate by treatment (group) interaction. If the assumption is tenable this means the interaction will *not* be significant. Therefore, what one does in SPSS MANOVA is to set up an effect involving the interaction (for one covariate), and then test whether this effect is significant. If the effect is significant, it means the assumption is not tenable.

For more than one covariate, as in the present case, there is an interaction term for each covariate. The effects are lumped together and then we test whether the combined interactions are significant. Before we give a few examples, note that BY is the keyword used by SPSS to denote an interaction, and + is used to lump effects together.

We show the control lines for testing the homogeneity assumption for two covariates and for three covariates. Denote the dependent variable by y , the covariates by x_1 and x_2 and the grouping variable by gp . The control lines are

```
ANALYSIS = Y /
DESIGN = X1+X2,GP,X1 BY GP+X2 BY GP /
```

Now, suppose there were three covariates. Then the control lines will be:

```
ANALYSIS = Y /
DESIGN = X1+X2+X3,GP,X1 BY GP+X2 BY GP+X3 BY GP /
```

The control lines for running the ANCOVA on the Sesame Street data with the covariates of PRENUMB and PRERELAT are given in Table 7.7. In Table 7.8 we present selected output from the SPSS analysis of covariance.

7.14 SUMMARY

1. In analysis of covariance a linear relationship is assumed between the dependent variable and the covariate(s).

2. ANCOVA is directly related to the two basic objectives in experimental design of (a) eliminating systematic bias and (b) reduction of error variance.

While ANCOVA does not eliminate bias, it can reduce bias. The use of several covariates with low intercorrelations will substantially reduce error variance.

3. Limit the number of covariates (C) so that

$$\frac{C + (J - 1)}{N} < .10$$

where J is the number of groups and N is total sample size.

4. A numerical example is given to show the intimate relationship between ANCOVA and the results for ANOVA on the same data.

TABLE 7.9

Placeholder for T0709 from p. 337 of previous edition.

5. Measurement error on the covariate causes loss of power in randomized designs, and can lead to seriously biased treatment effects in non-randomized designs.

6. In examining printout from the statistical packages, first make two checks to determine whether covariance is appropriate: (1) check that there is a linear relationship between the covariate and the dependent variable and (2) check that the regression slopes are equal. If *either* of these is not true, then covariance is not appropriate. In particular, if (2) is not true then the Johnson–Neyman technique should be considered.

7. Several cautions are given concerning the use of analysis of covariance with intact groups.

8. Three ways of analyzing a k group pretest-posttest design are: ANOVA on the difference scores, analysis of covariance, and a two way repeated measures ANOVA. Articles by Huck and McLean (1975) and by Jennings (1988) show that ANCOVA is generally the preferred method of analysis.

9. Although the Johnson–Neyman technique is preferred when the slopes are not equal, it is still not available on SAS or SPSS. *A violation of the equal slopes assumption means there is a group by covariate interaction effect.* Because of this, we illustrated, in Section 7.12, use of a two way ANOVA to get at the nature of this interaction.

10. We showed how ANCOVA can be done using multiple regression. By dummy coding group membership and appropriate multiplication we obtained both the test for homogeneity of regression slopes and the ANCOVA.

EXERCISES

1. A social psychological study by Novince (1977) examined the effect of behavioral rehearsal, and behavioral rehearsal plus cognitive restructuring (combination treatment) on reducing anxiety and facilitating social skills for female college freshmen. The 33 subjects were randomly assigned (11 each) to either BH, a control group (group 2), or BH + CR. The subjects were pretested and posttested on several variables. The scores for the avoidance variable are given as follows:

BEHAVIORAL REHEARSAL		CONTROL		BEHAVIORAL REHEARSAL + COGNITIVE RESTRUCTURING	
Avoid	Preavoid	Avoid	Preavoid	Avoid	Preavoid
91	70	107	115	121	96
107	121	76	77	140	120
121	89	116	111	148	130
86	80	126	121	147	145
137	123	104	105	139	122
138	112	96	97	121	119
133	126	127	132	141	104
127	121	99	98	143	121
114	80	94	85	120	80
118	101	92	82	140	121
114	112	128	112	95	92

Table 7.9 shows selected printout from an ANCOVA on SPSS for Windows 7.5 (top two thirds of printout).

- (a) Is ANCOVA appropriate for this data? Explain.
 - (b) If ANCOVA is appropriate, then do we reject the null hypothesis of equal adjusted population means at the .05 level?
 - (c) The bottom portion of the printout shows the results from an ANOVA on just avoidance. Note that the error term is 280.07. The error term for the ANCOVA is 111.36. How are the two error terms fundamentally related?
2. (a) Run an ANOVA on the difference scores for the data in exercise 1.
(b) Compare the error term for that analysis vs the error term for the ANCOVA on the same data. Relate these results to the discussion in Section 7.10.
 3. This question relates the use of a pretest as covariate to experimental design considerations. Suppose in a counseling study eight subjects were randomly assigned to each of three groups. The subjects were pretested and posttested on client satisfaction, which served as the dependent variable.
 - (a) What is the main reason for using the pretest here as a covariate?
 - (b) In what other way might the covariate be useful?
 - (c) What effect would the possibility of pretest sensitization have on your decision to use a pretest in this study?
 4. An analysis of variance is run on three intact groups and a significant difference is found at the .05 level. The pattern of means is

	GP 1	GP 2	GP 3
COVARIATE	120	100	110
DEP. VAR.	70	60	65

A few days later the investigator, after talking to a colleague, runs an ANCOVA on this data and no longer finds significance at the .05 level. The correlation between the dependent variable and the covariate is .61 and the homogeneity of regression slopes assumption is found to be tenable. Explain what has happened here, and relate this to the discussion in section 7.3.

5. A study by Huck and Bounds (1972) examined whether the grade assigned an essay test is influenced by handwriting neatness. They hypothesized that an interaction effect would occur, with graders who have neat handwriting lowering the essay grade while graders with messy handwriting will not lower the grade. Students in an Educational Measurement class at the University of Tennessee served as subjects. Sixteen were classified as having neat handwriting and 18 were classified as messy handwriters. Each of these 34 subjects received two one page essays. A person with average handwriting neatness copied the first (better) essay. The second essay was copied by two people, one having neat handwriting and one having messy handwriting. Each subject was to grade each of the two essays on a scale from 0 to 20. Within the neat handwriters, half of them were randomly assigned to receive a neatly written essay to grade and the other half a messy essay. The same was done for the messy handwriters who were acting as graders. The grade assigned to essay 1 served as the covariate in this study. Means and adjusted means are given below for groups:

		Neat Essay			Messy Essay
Neat Writer	Essay 1	$\bar{x} = 14.75$	Essay 1	$\bar{x} = 15.00$	
	Essay 2	$\bar{x} = 13.00$	Essay 2	$\bar{x} = 9.75$	
		$sd = 2.51$		$sd = 3.62$	
	Adj. Mean	$= 13.35$	Adj. Mean	$= 9.98$	
Messy Writer	Essay 1	$\bar{x} = 16.33$	Essay 1	$\bar{x} = 15.78$	
	Essay 2	$\bar{x} = 12.11$	Essay 2	$\bar{x} = 12.44$	
		$sd = 3.14$		$sd = 2.07$	
	Adj. Mean	$= 11.70$	Adj. Mean	$= 12.30$	

The following is from their RESULTS section (Huck & Bounds, 1972):

Prior to using analysis of covariance, the researchers tested the assumption of homogeneous within-group regression coefficients, Since this preliminary test proved to be nonsignificant ($F = 1.76, p > .10$), it was appropriate to use the conventional covariance analysis.

Results of the 2×2 analysis of covariance revealed that neither main effect was significant. However, an interaction between the legibility of the essay and the handwriting neatness of the graders was found to be significant ($F = 4.49, p < .05$). To locate the precise nature of this interaction, tests of simple main effects (Kirk, 1968, p. 481) were used to compare the two treatment conditions, first for graders with neat handwriting and then a second time for graders with messy handwriting. Results indicated that neat writers gave higher grades to the neat essay than to the messy essay ($F = 6.13, p < .05$), but that messy handwriters did not differentiate significantly between the two essays. (pp. 281–82)

- (a) From what is mentioned in the above RESULTS section, can we be confident that analysis of covariance is appropriate? Explain.
 - (b) What is the main reason for using analysis of covariance in this study?
 - (c) Should the investigators have been concerned about the homogeneity of cell population variances in this study? Why, or why not?
 - (d) Estimate the effect size for the interaction effect (see Section 4.6). Is it large or fairly large? Relate this to the sample size in the study and the significance that was found for the interaction effect.
6. Determine whether ANCOVA is appropriate for the HEADACHE data, using UNCOMF as the dependent variable and the PREUNCOMF as the covariate. What checks did you make?
 7. What is the main reason for using ANCOVA in a randomized study?
 8. Cochran, in his 1957 review article on ANCOVA, made the statement that ANCOVA will not be useful when the correlation between the dependent variable and the covariate is less than .3 in absolute value. Why did he say this?

Hierarchical Linear Modeling

Written by Dr. Natasha Beretvas

CONTENTS

- 8.1 Introduction
- 8.2 Problems Using Single-Level Analyses of Multilevel Data
- 8.3 Formulation of the Multilevel Model
- 8.4 Two-Level Model—General Formulation
- 8.5 HLM6 Software
- 8.6 Two-Level Example—Student and Classroom Data
- 8.7 HLM Software Output
- 8.8 Adding Level One Predictors to the HLM
- 8.9 Addition of a Level Two Predictor to a Two Level HLM
- 8.10 Evaluating the Efficacy of a Treatment
- 8.11 Final Comments on HLM

8.1 INTRODUCTION

In the social sciences, nested data structures are very common. As Burstein noted, “Most of what goes on in education occurs within some group context” (1980). Nested data (which yields correlated observations) occurs whenever subjects are clustered together in groups as is frequently found in social science research. For example, students in the same school will be more alike than students from a different school thereby implying some non-independence. Responses of patients to counseling for those patients clustered together in therapy groups will depend to some extent on the patient’s group’s dynamics resulting in a within-therapy group dependency (Kreft & deLeeuw, 1998). Yet *one of the assumptions made in many of the statistical techniques (including regression, ANOVA, etc.) used in the social sciences is that the observations are independent.*

Kenny and Judd noted that while non-independence is commonly treated as a nuisance, there are still “many occasions when nonindependence is the substantive problem that we are trying to understand in psychological research” (1986, p. 431). The authors refer to researchers interested in studying social interaction. Kenny and Judd note that social interaction by definition implies non-independence. If a researcher is interested in studying social interaction, or even a plethora of other social psychology constructs, the non-independence is not so much a statistical problem to be surmounted as a focus of interest.

Additional examples of dependent data can be found for employees working together in organizations, and even citizens within nations. These scenarios, as well as students nested within schools and patients within therapy groups, provide examples of two-level designs. The first level comprises the units that are grouped together at the second level. For instance, students (level one) would be considered as nested within schools (level two), and patients (level one) are nested within counseling groups (level two).

Examples of this nestedness of clustering does not always involve only two levels. A commonly encountered three-level design found in educational research involves students (level one) nested within classrooms (level two), clustered within schools (level three). Individuals (level one) are “nested” within families (level two) that are clustered in neighborhoods (level three). Patients (level one) are frequently counseled in groups (level two) that are clustered within counseling centers (level three). There is an endless list of such groupings. When data are clustered in these ways, use of multilevel modeling should be considered.

In the late 1970s, estimation techniques and programs were developed to facilitate use of multilevel modeling (Raudenbush & Bryk, 2002; Arnold, 1992). Before this time, researchers would tend to use single-level regression models to investigate relationships between relevant variables describing the different levels despite the violation of the assumption of independence. This would be problematic for a variety of reasons.

8.2 PROBLEMS USING SINGLE-LEVEL ANALYSES OF MULTILEVEL DATA

A researcher might be interested in the relationship between students’ test scores and characteristics of the schools that they attended. The dataset might consist of student and school descriptors from students’ who were randomly selected from a random selection of schools. When investigating the question of interest, a researcher choosing to ignore the inherent dependency in his or her data would have two analytical choices (other than the use of multilevel modeling). The researcher could aggregate the student data to the school level and use school data as the level of analysis. This would mean that the outcome in a single-level regression might

have been the school's average student score, with predictors consisting of school descriptors and average school characteristics summarized across students within each school. One of the primary problems with such an analysis is that valuable information is lost concerning variability of students' scores *within* schools, statistical power is decreased and the ecological validity of the inferences has been compromised (Hox, 2002; Kreft & de Leeuw, 1998).

Alternatively, the researcher could disaggregate the student- and school-level data. This modeling would have involved using students as the unit of analysis and ignoring the non-independence of students' scores within each school. In the single-level regression that would be used with disaggregated data, the outcome would be the student's test score with predictors including student and school characteristics. The problem in this analysis is that values for school descriptors would be the same across students within the same school. Using this disaggregated data, and thus ignoring the non-independence of the students' scores within each school, artificially deflates the estimated variability of the school descriptor. This would then affect the validity of the statistical significance test of the relationship between the student outcome and the school descriptor and inflate the associated Type I error rate. The stronger the relationship between students' scores for students within each school, the worse the impact on the Type I error rate.

There is a measure of the degree of dependence between individuals that is called the intra-class correlation (ICC). The more that characteristics of the context (say, school) in which an individual (student) finds himself have an effect on the outcome of interest, the stronger will be the ICC. In other words the more related to the outcome are the experiences of individuals within each grouping, the stronger will be the ICC (Kreft & de Leeuw, 1998). For two-level datasets (in which individuals there is only one level of grouping), the ICC can be interpreted as the proportion of the total variance in the outcome that occurs between the groups (as opposed to within the groups).

Snijders and Bosker (1999, p. 151) indicate that

"In most social science research, the intraclass correlation ranges between 0 and .4, and often narrower bounds can be identified."

Even an ICC that is slightly larger than zero can have a dramatic effect on Type I error rates as can be seen in the table taken from Scariano and Davenport (1987) on the following page.

Note from the table that for an ICC of only .01, with 3 groups and 30 subjects per group, the actual alpha is inflated to .0985 for a one way ANOVA. For a 3 group, $n = 30$ scenario in which $ICC = .10$, the actual alpha is .4917!

Fortunately, researchers do not have to choose between the loss of information associated with aggregation of dependent data nor the inflated Type I error rates associated with disaggregated data. Thus, instead of choosing a level at which to con-

Actual Type I Error Rates for Correlated Observations in a One Way ANOVA (Nominal $\alpha = .05$)

<i>Intraclass Correlation (ICC)</i>						
<i>m</i>	<i>n</i>	.00	.01	.10	.30	.50
2	3	.0500	.0522	.0740	.1402	.2374
	10	.0500	.0606	.1654	.3729	.5344
	30	.0500	.0848	.3402	.5928	.7205
	100	.0500	.1658	.5716	.7662	.8446
3	3	.0500	.0529	.0837	.1866	.3430
	10	.0500	.0641	.2227	.5379	.7397
	30	.0500	.0985	.4917	.7999	.9049
	100	.0500	.2236	.7791	.9333	.9705
5	3	.0500	.0540	.0997	.2684	.5149
	10	.0500	.0692	.3151	.7446	.9175
	30	.0500	.1192	.6908	.9506	.9888
	100	.0500	.3147	.9397	.9945	.9989

m—number of groups
n—number of observations per group

duct analyses of clustered or hierarchical data, researchers can instead use the technique called “multilevel modeling.” This chapter will provide an introduction to some of the simpler multilevel models. There are several excellent multilevel modeling texts available (Raudenbush & Bryk, 2002; Hox, 2002; Snijders & Bosker, 1999; Kreft & de Leeuw, 1998) that will provide the interested reader additional details as well as discussion of more advanced topics in multilevel modeling.

Several terms are used to describe essentially the same family of multilevel models including: multilevel modeling, hierarchical linear modeling, (co)variance component models, multilevel linear models, random-effects or mixed-effects models and random coefficient regression models, among others (Raudenbush & Bryk, 2002; Arnold, 1992). I will use “multilevel modeling” and “hierarchical linear modeling” in this introduction as they seem to provide the most comprehensible terms.

In this chapter, formulation of the multilevel model will first be introduced. This will be followed with an example of a two-level model. This example, which involves students within classes, we will first consider what is called an unconditional model (no predictors at either level). Then we consider adding predictors at level 1 and then a predictor at level 2. After this example we consider evaluating the efficacy of treatments on some dependent variable, and compare the HLM6 analysis to an SPSS analysis of the same data. In conclusion, we offer some final comments on HLM.

8.3 FORMULATION OF THE MULTILEVEL MODEL

There are two algebraic formulations possible for the hierarchical linear model (HLM). The set of equations for each level can be represented separately (while indexing the appropriate clusters), or alternatively, each level's equations can be combined to provide a single equation. The multiple levels equations formulation (Raudenbush & Bryk, 1992; 2002) seems to be the easiest to comprehend for a neophyte HLM user in that it simplifies the variance component assignment and clearly distinguishes the levels. This formulation also is the one that is implemented in the multilevel software HLM (Raudenbush, Bryk, Cheong, & Congdon, 2000). Because the HLM software will be used to demonstrate estimation of HLM parameters in this chapter, the multiple levels' formulation will be used.

8.4 TWO-LEVEL MODEL—GENERAL FORMULATION

Before presenting the general formulation of the two-level model, some terminology will first be explained. Raudenbush and Bryk (2002) distinguish between unconditional and conditional models. The unconditional model is one in which no predictors (at any of the levels) are included. A conditional model includes at least one predictor at any of the levels.

Multilevel modeling permits the estimation of fixed and random effects whereas ordinary least-squares (OLS) regression includes only fixed effects. For this reason, it is important to distinguish between fixed and random effects. If a researcher is interested in comparing two methods of counseling, for example, then the researcher would not be interested in generalizing beyond those two methods. The inferences would be "fixed" or limited to the two methods under consideration. Thus counseling method would be treated as a fixed factor. Similarly, if three diets (Atkins, South Beach, and Weight Watchers, for instance) were to be compared, then the diets were not randomly chosen from some population of diets, thus once again diets would be a fixed factor.

On the other hand, consider two situations in which a factor would be considered random. A researcher might be interested in comparing three specific teaching methods (fixed factor) used across schools in nine different random schools in some metropolitan area. The researcher would wish to generalize inferences about the teaching methods' effects to the population of schools in this area. Thus, here, schools is a random factor and teaching method effects would be modeled as randomly varying across schools. As a second example, consider the design in which patients are clustered together in therapy groups. Although a researcher would be interested in limiting her inferences to the specific counseling methods involved (fixed effect), she might want to generalize the inferences beyond the particular therapy groups involved. Thus groups would be considered a random factor and

counseling method effects modeled as randomly varying across groups. For further discussion of fixed and random effects, the interested readers should look at Kreft and de Leeuw's discussion (1998).

This two-level example will involve investigating the relationship between students' scores on a Mathematics achievement test in the 12th grade (*Math_12*) and a measure of the student's interest in mathematics (*IIM*). For students in a certain classroom, a simple one-level regression model could be tested:

$$Y_i = \beta_0 + \beta_1 X_i + r_i \quad (1)$$

where Y_i is student i 's grade 12 Math score, X_i is student i 's *IIM* score, β_1 is the slope coefficient representing the relationship between *Math_12* and *IIM*, and β_0 is the intercept representing the average *Math_12* score for students in the class's sample given a score of zero on X_i . The value of β_1 indicates the expected change in *Math_12* given a one unit increase in *IIM* score. The r_i represents the "residual" or deviation of student i 's *Math_12* score from that predicted given the values of β_0 , the student's X_i , and β_1 . It is assumed that r_i is normally distributed with a mean of zero and a variance of σ^2 , or $r_i \sim N(0, \sigma^2)$.

A brief note should be made about centering the values of a predictor. As mentioned above, the intercept, β_0 , represents the value predicted for the outcome, Y_i , given that X_i is zero. It is important to ensure that a value of zero for X_i is meaningful. Interval-scaled variables are frequently scaled so that they are "centered" around their mean. To center the *IIM* scores, they would need to be transformed so that student i 's value on X_i was the deviation of student i 's *IIM* score from the sample mean of the *IIM* scores. If this centered predictor were used instead of the original raw *IIM* score predictor, then the intercept β_0 would be interpreted as the predicted *Math_12* score for a student with an average *IIM* score.

A regression equation just like Equation 1 might be constructed for students in a second classroom. The relationship between *Math_12* and *IIM*, however, might differ slightly for the second classroom. Similarly, the coefficients in Equation 1 might be slightly different for other classrooms also. The researcher might be interested in understanding the source of the differences in the classrooms' intercepts and slopes. For example, the researcher might want to investigate whether there might be some classroom characteristic that lessens or overcomes the relationship between a student's interest in mathematics (*IIM*) and their performance on the math test (*Math_12*). To investigate this question, the researcher might obtain a random sample of several classrooms to gather students' *Math_12* and *IIM* scores as well as measures of classroom descriptors. Now regression equation 1 could be calculated for each classroom j such that:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \quad (2)$$

where the estimates for classroom j of the intercept, β_{0j} , and slope, β_{1j} might differ for each classroom. For each classroom's set of residuals, r_{ij} , it is assumed that their variances are homogeneous across classrooms, where $r_{ij} \sim N(0, \sigma^2)$.

The researcher would (hopefully) realize that given a large enough sample of classrooms' data, multilevel modeling could be used for this analysis. Math scores of students within the same classroom are likely more similar to each other than to scores of students in other classrooms. This dependency needs to be modeled appropriately. This brings us to the multiple sets of equations formulation of the HLM.

If multilevel modeling were to be used in the current example, then students are nested within classrooms. The higher level of grouping or clustering is associated with a higher value for the assigned HLM level. Thus, students will be modeled at level one and classrooms (within which students are "nested") at level two. The level one (student level) equation has already been presented (in Equation 2). The classroom level (level two) equations are used to represent how the lower level's regression coefficients might vary across classrooms. *The regression coefficients, β_{0j} and β_{1j} become response variables modeled as outcomes at the classroom level* (Raudenbush, 1984). Variation in classrooms' regression equations implies that the coefficients in these equations each might vary across classrooms. Variability in the intercept, β_{0j} , across classrooms would be represented as one of the level two equations by:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (3)$$

where β_{0j} is the intercept for classroom j , γ_{00} is the average intercept across classrooms (or, in other words, the average *Math_12* score across classrooms, controlling for *IIM* score) and u_{0j} is classroom j 's deviation from γ_{00} , where $u_{0j} \sim N(0, \tau_{00})$.

Variability in the relationship between *IIM* and *Math_12* (the slope coefficient) across classrooms is represented as a level two equation:

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (4)$$

where β_{1j} is the slope for classroom j , γ_{10} is the average slope across classrooms (or, in other words, the average measure of the relationship between *Math_12* and *IIM* scores across classrooms) and u_{1j} is classroom j 's deviation from γ_{10} , where $u_{1j} \sim N(0, \tau_{11})$. It is commonly assumed that the intercept and slope (β_{0j} and β_{1j}) are bivariate normally distributed with covariance τ_{01} (Raudenbush & Bryk, 2002).

The two level two equations (Equations 3 and 4) are usually more succinctly presented as:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{cases} \quad (5)$$

In this two-level unconditional model (see Equations 2 and 5) there are three sources of random variability: the level one variability, r_{ij} , the level two (across classrooms) variability in the intercept, u_{0j} , and in the slope, u_{1j} . An estimate of the level one variability, σ^2 , is provided. Estimates of the level-two variance components, τ_{00} and τ_{11} , (describing the variability of u_{0j} and u_{1j} , respectively) can each be tested for statistical significance.

Testing the variability of the intercept across classrooms assesses whether the variability of classrooms' intercepts (as measured using the associated variance component, τ_{00}) differs from zero. If it is inferred that there is not a significant amount of variability in the intercept (or if it is hypothesized based on theory that the intercept should not vary across classrooms) then the random effects variability term, u_{0j} , can be taken out of Equation 3 (or Equation 5) and the intercept is then modeled as fixed.

If, on the other hand, it is inferred that there is a significant amount of variability in the intercept across classrooms, then variables describing classroom (level two) characteristics can be added to the model in Equation 3 (or equation for β_{0j} in Equation 5) to help explain that variability. (This will be demonstrated later in the chapter). If the classroom characteristics are found to sufficiently explain the remaining variability in the intercept, then they can remain in the modified level two equation for the intercept and the random effect term can be taken out. With only level two predictors in Equation 3, the intercept is considered to be modeled as "non-randomly varying" (Raudenbush & Bryk, 2002).

The variability in the slope coefficients can also be tested by inspecting the statistical significance of the slope's variance component, τ_{11} . If it is inferred that there is a significant amount of variability in the slopes, (implying that the relationship between *Math_12* and *IIM* scores differs across classrooms), then a classroom predictor could be added to help explain the variability of β_{1j} (in Equation 4 or 5). The addition of a level two predictor to the equation for the slope coefficient would be termed a "cross-level interaction" which is an interaction between variables describing different clustering levels (Hox, 2002). The variance component remaining (conditional upon including the level two predictor) can then be tested again to see if it sufficiently explained the random variability in slopes. With the addition of a predictor that does influence the relationship between the level one variable (here, *IIM*) and the outcome (*Math_12*), the remaining variability will be lowered as will be the associated variance component, τ_{11} . The values of the level two variance components (for the intercept and slope coefficients) can be compared with their values in the unconditional (no predictors) model to assess the proportion of (classroom) level two variability explained by the predictors that were added to the model in Equation 5. This, as well as addition of level one and level two predictors to the model will be demonstrated further in the next section.

Having discussed the formulation of the two-level HLM, use of the HLM software (version 6) will now be introduced and then demonstrated using a worked-ex-

ample. This example will be presented to demonstrate the process of HLM model-building involving addition of predictors to the two levels of equations, as well as interpretation of the parameter estimates presented in the HLM output.

8.5 HLM6 SOFTWARE

Raudenbush, Bryk, Cheong and Congdon's (2004) HLM software, version 6, for multilevel modeling provides a clear introduction for beginning multilevel modelers. In addition, it is possible for students to obtain a free-ware copy of the program for simple multilevel analyses. This provides beginners with an easy way to evaluate for themselves whether they wish to purchase the entire program. (The website is WWW.SSICENTRAL.COM)

The SS is an abbreviation for Scientific Software, which produces and distributes the HLM software. When you get to this site, click on HLM. You will get a dropdown menu, at which point click on free downloads.

The datasets being analyzed by HLM can be in any of the following formats: ASCII, SPSS, SAS portable, or SYSTAT. One of the complications of using HLM is that *separate* data files must be constructed for each level of clustering. For example, when investigating a two-level dataset, the user must construct a level one file as well as a level two file. These two files *must* be linked by a common id on both files. (This will be *TeachId* in the example we are about to use). **Data analysis via HLM involves four steps:**

1. **Construction of the data files.**
2. **Construction of the multivariate data matrix (MDM) file, using the data files.**
3. **Execution of analyses based on the MDM file.**
4. **Evaluation of the fitted model(s) based on a residual file.**

We will not deal with step 4 as this chapter is an introduction to HLM.

8.6 TWO-LEVEL EXAMPLE—STUDENT AND CLASSROOM DATA

The first step in using HLM to estimate a multilevel model is to construct the relevant datasets. As mentioned, for a two-level analysis, two data files are needed: one for each level. The level two ID variable (*TeachId*) in the current example must appear in both files.

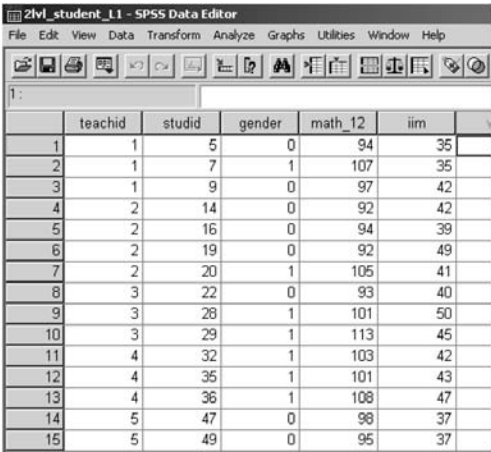
In this example, the researcher is interested in the relationship between scores on a 12th grade mathematics test (*Math_12*) and student and classroom character-

istics. The researcher has information about students’ gender and their individual scores on an interest in mathematics (*IIM*) inventory and on the outcome of interest (*Math_12*). Thus, *Math_12*, *IIM*, and *Gender* as well as the *TeachId* identifying the teacher/classroom for each student must appear in the level one dataset.

The researcher also has a measure of each classroom’s “resources” (*Resource*) that assesses the supplies (relevant to mathematics instruction) accessible to a classroom of students. Thus the level two dataset will contain *Resource* and *TeachId*. We will use SPSS data files.

Setting up the Datasets for HLM Analysis

The level one dataset contains the level two id (*TeachId*) as well as the relevant student-level descriptors (*IIM* and *Gender*) and outcome (*Math_12*). Another minor complication encountered when using HLM is that the data should be sorted by level two id and within level two id, by student id. A snapshot of the Level one dataset appears in Figure 8.1. The raw data files are given at the end of the chapter.

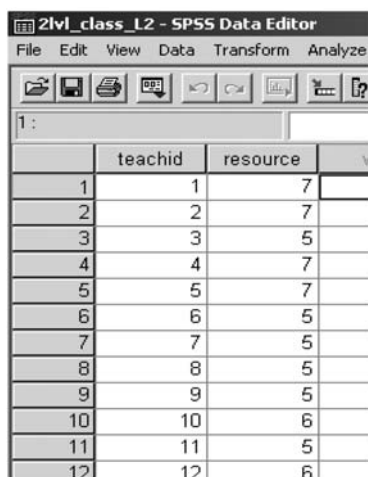


The image shows a screenshot of the SPSS Data Editor window titled "2lvl_student_L1 - SPSS Data Editor". The window displays a dataset with 15 rows and 6 columns. The columns are labeled: teachid, studid, gender, math_12, iim, and an unlabeled column with a dropdown arrow. The data is organized by teachid (1, 2, 3, 4, 5) and then by studid within each teachid. The gender column contains 0 or 1, math_12 contains integer scores, and iim contains integer scores.

	teachid	studid	gender	math_12	iim	
1	1	5	0	94	35	
2	1	7	1	107	35	
3	1	9	0	97	42	
4	2	14	0	92	42	
5	2	16	0	94	39	
6	2	19	0	92	49	
7	2	20	1	105	41	
8	3	22	0	93	40	
9	3	28	1	101	50	
10	3	29	1	113	45	
11	4	32	1	103	42	
12	4	35	1	101	43	
13	4	36	1	108	47	
14	5	47	0	98	37	
15	5	49	0	95	37	

FIGURE 8.1 Two-level model–student level SPSS dataset.

As can be seen in Figure 8.1, the dataset is set up to mimic the clustering inherent in the data. Students are “nested” within classrooms that are identified using the variable *TeachId*. The first classroom (*TeachId* = 1) provides student-level information on three students (students 5, 7 and 9). The second classroom provides data for four students (14, 16, 19 and 20), and so on. The level two dataset appears in Figure 8.2 below.



The screenshot shows the SPSS Data Editor window titled "2lvl_class_L2 - SPSS Data Editor". The menu bar includes File, Edit, View, Data, Transform, and Analyze. Below the menu bar is a toolbar with icons for opening, saving, printing, and other functions. The data grid shows 12 rows of data with two columns: "teachid" and "resource". The "teachid" column contains values from 1 to 12, and the "resource" column contains values from 5 to 7. The rows are numbered 1 through 12 in the first column.

	teachid	resource
1	1	7
2	2	7
3	3	5
4	4	7
5	5	7
6	6	5
7	7	5
8	8	5
9	9	5
10	10	6
11	11	5
12	12	6

FIGURE 8.2 Two-level model—classroom level spss dataset.

In the level two dataset, the classroom information (here, the *TeachId* and the classroom’s score on the *Resource* measure) are listed. Note that the *TeachId* values are ordered in both the level one and level two files as required by HLM software.

Setting up the MDM File for HLM Analysis

Before using HLM, the user needs to first construct what is called the “multivariate data matrix” or MDM file that sets up the datasets (regardless of their original format) into a format that can be used more efficiently when running the HLM program. (Note that in prior versions of HLM, an SSM file was constructed instead of an MDM file). Once the datasets are set up in SPSS (or other relevant statistical software programs) the following steps are taken to set up the MDM file.

1. Once the HLM program is opened, click on FILE, scroll down to “MAKE NEW MDM FILE” and request STAT PACKAGE INPUT as shown in the following screen (Figure 8.3).

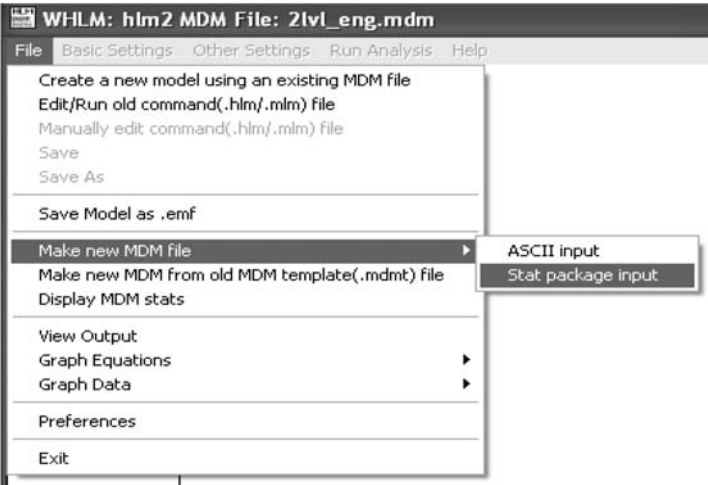


FIGURE 8.3 First HLM window for building MDM file.

2. You must then identify the kind of modeling to be used from the window displayed below. Choose HLM2 for this two-level example and click on OK.

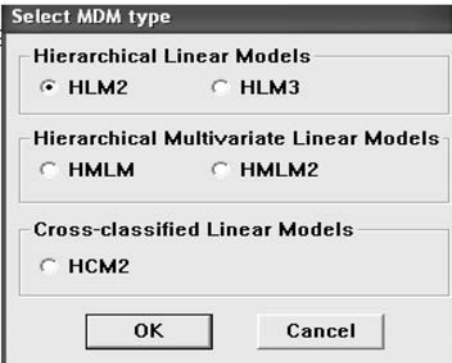


FIGURE 8.4 Second HLM window for building MDM file.

3. After clicking on HLM2, the “Make MDM—HLM2” HLM window appearing below in Figure 8.5 will appear. Fill in a filename for the MDM file (under MDM File Name) being sure to include “.MDM” as the suffix. Given SPSS datasets are being analyzed, make sure to change INPUT FILE TYPE to SPSS/WINDOWS before attempting to find the relevant level one and two data files. Because the first multilevel example involves students nested within classrooms, be sure to click on “persons within groups” instead of “measures within persons”. Select the level one data file by clicking on BROWSE under LEVEL-1 SPECIFICATION and finding the relevant file (here, called 2lvl_student_L1.SAV). Note that the level one and level two SPSS files that are going to be used in the analysis should not be open in SPSS when the user is constructing the MDM file.

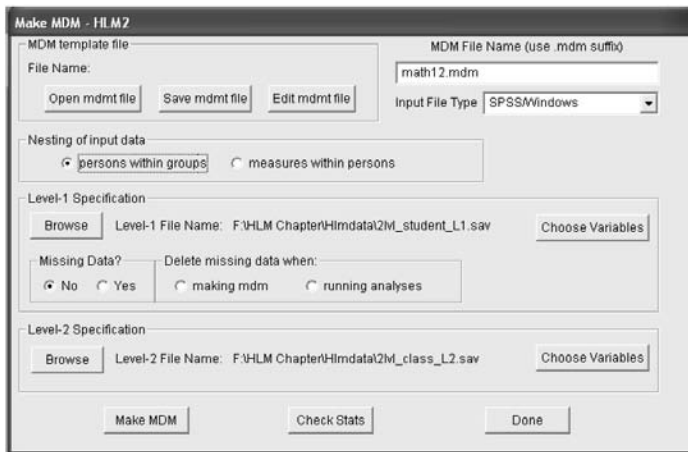


FIGURE 8.5 Third HLM window for building MDM file.

4. Click on CHOOSE VARIABLES and select the level two id (*TeachId*) that links the level one and two files as well as the relevant level one variables (*Gender*, *Math_12*, and *IIM* in the current example). Figure 8.6 displays this screen. In both Figures 8.6 and 8.7 it should read “in MDM” (since we using version 6 of HLM).

5. Follow the same procedure to identify the relevant level two file for use in the MDM by clicking on BROWSE and finding the level two .SAV file (here, the 2lvl_class_L2.SAV file). Again, click on CHOOSE VARIABLES and identify the level two id (*TeachId*) and the level two variables of interest (just *Resource* in the current example). The level two CHOOSE VARIABLE screen appears in Figure 8.7 on the following page.

The screenshot shows a dialog box titled "Choose variables - HLM2". It contains a list of variables on the left and a grid of checkboxes on the right. The variables listed are TEACHID, STUDID, GENDER, MATH_12, and IIM. The checkboxes for "ID" and "in SSM" are visible for each variable. The "ID" checkbox for TEACHID is checked, while the "in SSM" checkbox is unchecked. The "ID" checkbox for STUDID is unchecked, while the "in SSM" checkbox is checked. The "ID" checkbox for GENDER is unchecked, while the "in SSM" checkbox is checked. The "ID" checkbox for MATH_12 is unchecked, while the "in SSM" checkbox is checked. The "ID" checkbox for IIM is unchecked, while the "in SSM" checkbox is checked. At the bottom of the dialog box, there is a "Page 1 of 1" indicator, a navigation bar with left and right arrows, and "OK" and "Cancel" buttons.

FIGURE 8.6 Setting up an MDM File—Choosing Variables at Level One

The screenshot shows a dialog box titled "Choose variables - HLM2". It contains a list of variables on the left and a grid of checkboxes on the right. The variables listed are TEACHID, RESOURCE, and several empty rows. The checkboxes for "ID" and "in SSM" are visible for each variable. The "ID" checkbox for TEACHID is checked, while the "in SSM" checkbox is unchecked. The "ID" checkbox for RESOURCE is unchecked, while the "in SSM" checkbox is checked. At the bottom of the dialog box, there is a "Page 1 of 1" indicator, a navigation bar with left and right arrows, and "OK" and "Cancel" buttons.

FIGURE 8.7 Setting up an MDM file—choosing variables at level two.

6. Next, you need to click on “Save mdmt file” (to save the MDM template file) and provide a name for the response (.MDMT) file.

7. Click on “Make MDM” to ensure that the data has been input correctly. A MS-DOS window will briefly appear (after clicking on MAKE MDM) ending in a count of the number of level two and level one units. If there seems to be a disparity between the group and within-group sample sizes, make certain that the original data files are sorted by the level two id.

8. Before you can exit the MAKE MDM window, you must also click on CHECK STATS. Once this is done, you can click on DONE to be brought to the HLM window that allows you to build the model to be estimated.

The Two-Level Unconditional Model

The unconditional model (including no predictors) is the model typically estimated first when estimating multilevel models. Estimation of the unconditional model provides estimates of the partitioning of the variability at each level. In the current example, this means that the variability between students can be estimated and the variability can be estimated between classrooms. If there is not a substantial amount of variability between classrooms, then this additional level of clustering might not be needed.

At level one, in the unconditional model, the outcome (*Math_12*) for student i in classroom j is modeled only as a function of classroom j 's intercept (i.e. average *Math_12* score) and the student's residual:

$$\text{Math_12}_{ij} = \beta_{0j} + r_{ij} \quad (6)$$

At level two, classroom j 's intercept is modeled to be a function of the average intercept (*Math_12* score) across classrooms and a classroom residual:

$$\beta_{0j} = \gamma_{00} + u_{0j} \quad (7)$$

HLM's presentation of these equations is very similar to Equations 6 and 7 although it does not include the relevant i and j subscripts.

Estimating Parameters of the Two-Level Unconditional Model

Once the MDM file is built, the HLM window that you can use to build your model appears with the newly constructed MDM automatically loaded.

After the MDM is loaded, a blank formula screen appears with the list of level one variables appearing on the left-hand side of the screen. The steps necessary to build the unconditional two-level model are as follows:

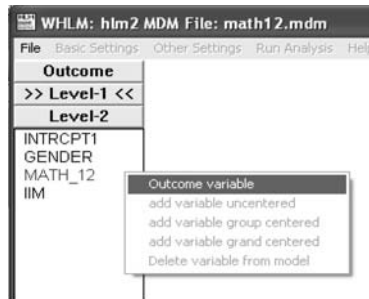


FIGURE 8.8 Selecting the outcome variable in HLM.

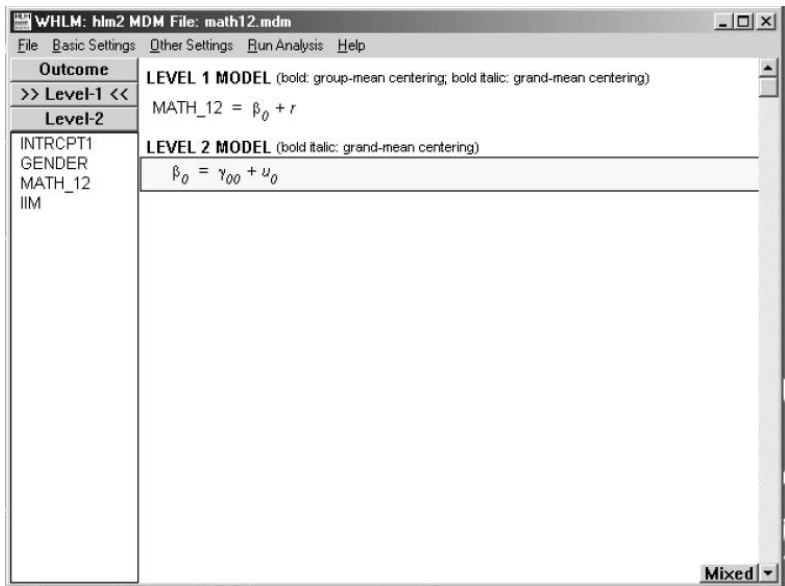


FIGURE 8.9 Unconditional model in HLM for two-level model.

1. Once the relevant MDM is loaded, the first thing a user must do is choose the relevant outcome variable (here, *Math_12*). Thus, click on *Math_12* and then OUTCOME VARIABLE as is shown in Figure 8.8.

For a two-level model, HLM automatically presents the two-level “unconditional model” with no predictors at levels one nor two as is shown in Figure 8.9.

Basic Model Specifications - HLM2

Distribution of Outcome Variable

☒ Normal (Continuous)

☐ Bernoulli (0 or 1)

☐ Poisson (constant exposure)

☐ Binomial (number of trials) None

☐ Poisson (variable exposure)

☐ Multinomial Number of categories

☐ Ordinal

☐ Over dispersion

Level-1 Residual File Level-2 Residual File

Title Unconditional two-level model

Output file name F:\HLM Chapter\Hlmdata\two_lev.out

Graph file name F:\HLM Chapter\Hlmdata\grapheq.geq

Cancel OK

FIGURE 8.10 HLM basic model specification model.

If you wish to run the model (without saving it) and examine the output, click on RUN ANALYSIS. When you click on RUN ANALYSIS, the program will respond that the model has not been saved; just click on RUN THE MODEL SHOWN (wait several seconds). Then click on FILE and scroll and click on VIEW OUTPUT. By doing this you can skip steps 2 through 5 below.

2. Click on BASIC SETTINGS to change the output file name from the default HLM2.TXT to something meaningful (like TWO_LEV.OUT as demonstrated below). It also helps to change the Title of the model from “no title” to something like “Unconditional two-level model” as this will appear on every page of the output.

```
Computing . . . , please wait
Starting values computed. Iterations begun.
Should you wish to terminate the iterations prior to convergence, enter cntl-c
The value of the likelihood function at iteration 1 = -4.725440E+002
The value of the likelihood function at iteration 2 = -4.725364E+002
The value of the likelihood function at iteration 3 = -4.725340E+002
The value of the likelihood function at iteration 4 = -4.725332E+002
The value of the likelihood function at iteration 5 = -4.725328E+002
The value of the likelihood function at iteration 6 = -4.725328E+002
```

FIGURE 8.11 HLM DOS window presenting iterations while HLM is running.

For details about the remaining options, the reader can refer to the HLM manual (Raudenbush, Bryk, Cheong, & Congdon, 2004). Click OK.

3. Save the model by clicking on FILE, then SAVE AS and typing in the model's filename.

4. Click on RUN ANALYSIS. Once the solution has converged, the MS-DOS window displaying the iterations (see below) will close and bring you back to the HLM model screen. (Based on HLM's defaults, if more than 100 iterations are needed, the user will be prompted whether the program should be allowed to iterate until convergence. For the current dataset, only six iterations were needed until the convergence criteria were met.

5. You can view the HLM output by clicking on FILE and then VIEW OUTPUT.

8.7 HLM SOFTWARE OUTPUT

The output containing the model's parameter estimates can be viewed if the user clicks on File \Rightarrow View Output. The equations match the format of those presented in the original HLM window when the model was being built. This part of the output appears as follows:

Summary of the model specified (in equation format)

```
Level-1 Model
      Y = B0 + R
Level-2 Model
      B0 = G00 + U0
```

The listing of the equations' coefficients is useful when the user needs to interpret the later output. Following the listing of the equations, the iterations and starting estimates for the various parameters are listed. After the information about the last iteration needed for the model's estimation, the message "Iterations stopped due to small change in likelihood function" appears and the results that follow include final parameter estimates.

The first parameter estimate that appears is the variance, σ^2 , of students' *Math_12* scores within classrooms (assumed homogeneous across classrooms). The value for the current data set is 50.47. The only other level two variance component that is estimated (in this unconditional model) represents the variability of classrooms' intercepts, τ_{00} . The value of the τ_{00} estimate is 26.42 for the current example. Next, the reliability estimate of β_{0j} as an estimate of γ_{00} is provided and is .688 for the current data set. This indicates that the classrooms' intercept estimates tend to provide moderately reliable estimates of the overall intercept (see the HLM

manual (Raudenbush, Bryk, Cheong, & Congdon, 2000) and Raudenbush & Bryk's (2002) HLM text for more information about this form of reliability estimate).

In the output, there are two tables containing estimates of the relevant fixed effect(s). The second table lists the fixed effects estimates along with "robust standard errors." These should be used when summarizing fixed effects, however, if the standard errors in the two fixed effects' tables differ substantially, then the user might wish to re-consider the fit of some of the assumptions underlying the model being estimated. The table containing the fixed effect estimate with robust standard appears below:

Final estimation of fixed effects (with robust standard errors)						
Fixed Effect		Coefficient	Standard Error	T-ratio	Approx. df	P-value
For	INTRCPT1, B0					
	INTRCPT2, G00	98.043234	1.112063	88.163	29	0.000

The only fixed effect estimated in the two-level unconditional model is the intercept, γ_{00} (see Equation 7). The estimate of the average *Math_12* value across schools is 98.04 with a standard error of 1.11. This coefficient differs significantly from zero ($t(29) = 88.163, p < .0001$).

The next part of the output presents the estimates of the variance components. We have two variance components that are estimated, the variability within classrooms, σ^2 , and the variability between classrooms, τ_{00} . Values of these two components' estimates were presented earlier in the output (as mentioned above) but also appear in table summary appearing as follows in the HLM output:

Final estimation of variance components:						
Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	5.14032	26.42291	29	96.72024	0.000
level-1,	R	7.10441	50.47257			

The variance component estimates match those mentioned earlier. The value of the τ_{00} estimate can be tested against a value of zero using a test statistic that is assumed to follow a χ^2 distribution (Raudenbush & Bryk, 2002). The results indicate that we can infer that there is a statistically significant amount of variability in

Math_12 scores between classrooms ($\sigma^2(29) = 96.72, p < .0001$). This supports the two-level modeling of the clustering of students' *Math_12* scores within classrooms.

The estimates of the variance components can be combined to provide an additional descriptor of the possible nestedness of the data. The intraclass correlation provides a measure of the proportion of the variability in the outcomes that exists between units of one of the multilevel model's levels. Specifically, for the two-level model estimated here, the intraclass correlation provides a measure of the proportion of variability in *Math_12* between classrooms. The formula for the intraclass correlation for a two-level model is:

$$\rho_{ICC} = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \quad (8)$$

For the current data set, the intraclass correlation estimate is

$$\hat{\rho}_{ICC} = \frac{\hat{\tau}_{00}}{\hat{\tau}_{00} + \hat{\sigma}^2} = \frac{26.42}{26.42 + 50.47} = .34$$

which means that 34% of the variability in *Math_12* scores is estimated to lie between classrooms (and thus it can be inferred that about 66% lies *within* classrooms).

The last information appearing in the HLM output consists of the deviance statistic that can be used to compare the fit of a model to the data when comparing two models. (It should be noted that to use the Deviance statistic to compare models one model must be a simplified version of the other in that some of the parameters estimated in the more parameterized model are not estimated but are instead constrained to a certain value in the simplified model). For the current unconditional model estimated the deviance statistic's value is 945.07 with two covariance parameters estimated (σ^2 and τ_{00}).

Since a substantial amount of variability was found both within and among classrooms, student and classroom descriptors could be added to the model to explain some of this variability. We will start by adding two student predictors to the level one equation.

8.8 ADDING LEVEL ONE PREDICTORS TO THE HLM

The dataset contains two student descriptors including *Gender* and interest in mathematics (*IIM*) scores. The researcher was interested in first including *IIM* scores as a level one predictor of *Math_12* scores. To add a level one variable to a model using HLM software, the user must click on the relevant variable. When a

variable is clicked on, HLM prompts for the kind of centering that is requested for the variable. The choices include: add variable *uncentered*, add variable *group centered*, and add variable *grand centered*.

Centering

Before continuing with the description of the formulation of the model using HLM software, brief mention should be made of centering. It should be remembered that even in a simple, *single-level* regression model ($Y_i = \beta_0 + \beta_1 X_i + e_i$) including a predictor, X_i , the intercept represents the average value of the outcome, Y_i , for person i with a zero on X_i . Users of single-level regression can “center” their predictors to ensure that the intercept is meaningful. This centering can be done by transforming subjects’ scores on X_i so that X_i represents a person’s deviation from the sample’s mean on X_i . This would transform interpretation of the single-level regression equation’s intercept to be the average value of Y_i for someone at the (sample) mean on X_i .

Alternatively, the simple regression might model the relationship between a dichotomous predictor variable [representing whether a subject was in the placebo (zero dosage) group or a treatment (10mg dosage) group] and some measure of, say, anxiety. The predictor could be dummy-coded such that a value of zero was assigned for those in the placebo group with a value of one for those in the treatment group. This would mean that the intercept would represent the predicted anxiety level for a person who was in the placebo condition.

The importance of assigning a meaningful reference point for a value of zero for the predictors in single-level regression extends to the inclusion of interactions between predictors in the single-level model. The reason for this is that the interpretation of a main effect can be affected by the inclusion of an interaction between predictors (resulting in the model: $Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i * Z_i + e_i$). Specifically, if an interaction is modeled between, say, predictor variables X and Z , then the coefficient for the main effect of X represents the *effect of X given Z is zero*. Thus you want to ensure that a value of zero on Z is meaningful. Similarly, the main effect of Z would be interpreted (with the interaction of X and Z included in the model) as *the effect of Z given X is zero*.

The need for centering predictor variables extends beyond single-level regression equations to include multilevel modeling. In a two-level multilevel model, a choice of centering is available for any level-one predictor variables included in the level one equation. The level one equation depicted in Equation 2 ($Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij}$) represents a single level-one predictor, X_{ij} , added to the model to help explain variability in the outcome, Y_{ij} . As in a single-level regression equation, the intercept, β_{0j} , represents the predicted value of Y_{ij} for someone with $X_{ij} = 0$.

As in single-level regression, a score of zero on X_{ij} might be meaningful (as in the example in which membership in a placebo condition might be assigned a zero

on X_{ij} as compared with a value of one assigned to those in a treatment condition). However, sometimes, a value of zero on the untransformed scale of X_{ij} might be unrealistic. Raudenbush and Bryk (2002) use an example in which X_{ij} is a subject's SAT score for which feasible values only range from 200 to 800. In scenarios in which the value of zero on untransformed X_{ij} is not meaningful, a researcher should center his/her predictor variable.

Given a two-level model, there are two primary options (beyond not centering at all) for centering the level one predictor variable. One option involves centering the variable around the grand mean of the sample (as was described as an alternative for single-level regression), appropriately termed "grand-mean centering". This is accomplished by transforming the score on X_{ij} of subject i from group j (where, in the current example being demonstrated using HLM software, the grouping variable was "school") into the deviation of that score, X_{ij} , from the overall sample's mean score on X_{ij} (represented as $\bar{X}_{..}$). These transformed scores ($X_{ij} - \bar{X}_{..}$) are then used as the predictor of the outcome Y_{ij} in Equation 2. This means that the intercept term in Equation 2 represents the predicted value on Y_{ij} for someone with a value of zero on the predictor: $(X_{ij} - \bar{X}_{..})$. A subject with a value of zero on the predictor has an X_{ij} value equal to the grand mean: $\bar{X}_{..}$. Thus the intercept is the predicted value on Y_{ij} for someone at the grand mean on X_{ij} . This grand-mean centering results in the intercept being interpretable as the mean on Y_{ij} for group j adjusted by a function of the deviation of the group's mean from the grand mean (Raudenbush & Bryk, 2002).

In multilevel modeling, another alternative is available for centering a level one predictor variable. This alternative is termed "group-mean centering" and involves transforming the score, X_{ij} , of person i in group j into the deviation of that person's score from that (that person's) group j 's mean on X_{ij} : $(X_{ij} - \bar{X}_{.j})$. This modifies interpretation of the intercept, β_{0j} , so that it becomes the predicted value on Y_{ij} for someone with zero for $(X_{ij} - \bar{X}_{.j})$, or someone with a score that is the equivalent of group j 's mean on X_{ij} .

Several authors (including Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002) provide a detailed explanation for the correspondence between a model in which grand-mean centering is used and one in which variables are not centered. Essentially, when grand mean centering is used, a constant (the sample's mean on the relevant predictor) is subtracted from each case's value on the predictor. This means that the parameter estimates resulting from grand-mean centering can be linearly transformed to obtain the relevant uncentered variables' model's coefficients. This is not always the case when a variable has been group-mean centered. In group-mean centering, the mean of the case's group on the group-mean centered predictor is subtracted from the case's value on a predictor. Clearly, each group's mean will not be the same on the predictor and thus the same constant is not subtracted from each case's predictor value. The correspondence between a model

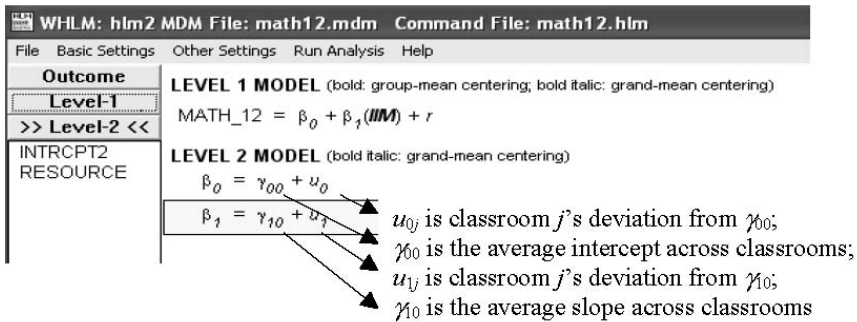


FIGURE 8.12 Adding a level one predictor to a two-level model in HLM.

with group-mean centered variables and models without centering or with grand-mean centering is not generally direct.

The reader should also be cautioned that, as in single-level modeling, choice of centering for predictors also impacts interpretation of main effects for variables when interactions that include that variable are modeled. This applies in multilevel modeling to same-level interactions between predictors as well as cross-level interactions in which, say, a level two predictor might be used to explain the relationship between a level one predictor and the outcome of interest.

Choice of grand-mean versus group-mean centering clearly impacts the interpretation of the intercept. However, as described in detail by Raudenbush and Bryk (2002), the choice of centering can also impact estimation of the level-two variances of the intercept and of the slope or coefficient of the predictor across groups (here, schools). This means that estimation of the variance in the u_{0j} s and the u_{1j} s (see Equation 5) will also be impacted by whether group-mean centering or grand-mean (and/or no centering) is used. As summarized by Raudenbush and Bryk: “be conscious of the choice of location for each level-1 predictor because it has implications for interpretation of β_{0j} , $\text{var}(\beta_{0j})$ and by implication, all of the covariances involving β_{0j} . In general, sensible choices of location depend on the purposes of the research. No single rule covers all cases. It is important, however, that the researcher carefully consider choices of location in light of those purposes; and it is vital to keep the location in mind while interpreting results” (Raudenbush & Bryk, 2002, p. 34). Several authors provide more detailed discussion of choice of centering than can be presented here (Snijders & Bosker, 1999; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002). The reader is strongly encouraged to refer to these texts to help understand centering in more detail.

In the example we are using to demonstrate use of HLM software, we will use grand mean centering for the *IIM* variable. *IIM* is added as a grand-mean centered

level one variable by clicking on the variable and requesting “add variable grand centered.” *In version 6 of HLM, the default is for a predictor’s effect to be modeled as fixed. (In version 5 of HLM, the default was for the effect to be random).* To model this effect as random, click on the level two equation for the coefficient of *IIM* and $\beta_1 = \gamma_{10}$ will become $\beta_1 = \gamma_{10} + u_1$ (see Figure 8.12). Again, the HLM model does not present the relevant *i* and *j* subscripts (see Figure 8.12).

Note that the regression coefficients in level 1 are response (dependent) variables in level 2. In this regard, the following from Kreft and DeLeeuw (1998, p. 2) is very important, “It is essential to realize that multilevel models involve a *statistical integration* of the different models specified at the levels of interest. The simplest integration takes place in the random coefficients model, where the first level regression coefficients are treated as random variables at the second level.”

The output appears as before although with additional parameters estimated given this second model includes an additional predictor. The fixed effect estimates will be presented and discussed first and then the random effects estimates. The user should be reminded that the first results that appear in HLM output are initial estimates. The user needs to look at the end of the output file to find the final estimates!

Only two fixed effects were modeled: the intercept, γ_{00} , and the slope, γ_{10} :

Final estimation of fixed effects (with robust standard errors)						
Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. df	P-value	
For INTRCPT1, B0 INTRCPT2, G00	98.768313	0.886268	111.443	29	0.000	
For IIM slope, B1 INTRCPT2, G10	0.899531	0.172204	5.224	29	0.000	

From the results (above), both parameter estimates differ significantly from zero. The intercept, γ_{00} , estimate is 98.77 ($t(29) = 111.44, p < .0001$) and the slope, γ_{10} , estimate is .90 ($t(29) = 5.22, p < .0001$). This means that the average *Math_12* score, controlling for *IIM*, is predicted to be 98.77. Here, due to the grand-mean centering of *IIM*, the “controlling for *IIM*” can be interpreted as: “for a student with the mean score on *IIM*.” The value of the slope coefficient estimate represents an estimate of the change in *Math_12* score predicted for a change of one in *IIM* score. Thus, these fixed effects coefficient estimates are interpreted very similarly to coefficients in OLS regression. Here, the higher a student’s *IIM* score, the higher will be their predicted *Math_12* score.

The output describing the random effects estimates appear at the end of the output and are as follows:

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	3.90385	15.24001	29	65.22134	0.000
IIM slope,	U1	0.68000	0.46239	29	48.11048	0.014
level-1,	R	5.10085	26.01870			

The level one variance explained by the addition of *IIM* to the model is seen in the reduction of the level one variance estimate, σ^2 , from a value of 50.47 in the unconditional model to a value of 26.02 in the current conditional model. In fact the proportion of the level one variance explained with the addition of *IIM* to the model is: $(50.47 - 26.02)/50.47 = .4844$ or 48.44%. In terms of the variability in the outcome among classrooms, there is still a significant amount of variability remaining in the intercept ($\hat{\tau}_{00} = 15.24$, $\chi^2(29) = 65.22$, $p < .0001$). It cannot be assumed that the average *Math_12* score controlling for *IIM* can be assumed constant across classrooms. There is also a significant amount of variability in the *IIM* slope coefficient across classrooms ($\hat{\tau}_{11} = .46$, $\chi^2(29) = 48.11$, $p < .05$). Thus it cannot be assumed that the relationship between *IIM* and *Math_12* can be assumed fixed across classrooms.

Additional random effects information appears in the output right after the information about the starting values and iterations required for convergence.

Tau			
INTRCPT1,	B0	15.24001	0.43587
IIM,	B1	0.43587	0.46239
Tau (as correlations)			
INTRCPT1,	B0	1.000	0.164
IIM,	B1	0.164	1.000

The first “Tau” (τ) matrix provides the estimates of the elements of the covariance matrix of level two random effects: $\begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}$ where τ_{00} is the variance of the intercept residuals u_{0j} , τ_{11} is the variance of the slope residuals, u_{1j} , and τ_{01} is the covariance between the random effects, u_{0j} and u_{1j} . The second Tau matrix is the correlation matrix corresponding to the first Tau matrix. It seems that there is not a strong correlation ($r = .164$) between the intercepts and the slopes.

The last lines in the output indicate that the deviance of this second model is 880.09 associated with four covariance parameters that are estimated (including σ^2 , τ_{00} , τ_{11} , τ_{01}). The difference in the deviances between the unconditional model and the current conditional model is assumed to follow a (large-sample) χ^2 distribution with degrees of freedom (DF) equal to the difference in the number of random effects parameters that are estimated in the two “nested” models. The difference in the deviances: $945.07 - 880.10 = 64.97$ can thus be tested against a χ^2 statistic with 2 DF. The statistical significance of the deviance difference indicates that the fit of the simpler (unconditional) model is significantly worse and thus the simpler model should be rejected.

**Adding a Second Level One Predictor
to the Level One Equation**

Because there still remains a substantial amount of variability in *Math_12* within classrooms, and since the researcher might hypothesize that there are gender differences in *Math_12* scores (controlling for *IIM*), a second level one predictor (*Gender*) will be added to the level one model. This is simply accomplished (in HLM software) by clicking on the relevant *Gender* variable. The variable *Gender* is coded with a zero for males and a one for females. The variable will be added as an uncentered predictor. Again, the default in HLM version 6 for adding a predictor is that it is to be modeled as a fixed effect. Click on the effect to change it so it is modeled as random and thus the level one equation to be estimated is:

$$Math_12_{ij} = \beta_{0j} + \beta_{1j}IIM_{ij} + \beta_{2j}Gender_{ij} + r_{ij} \tag{9}$$

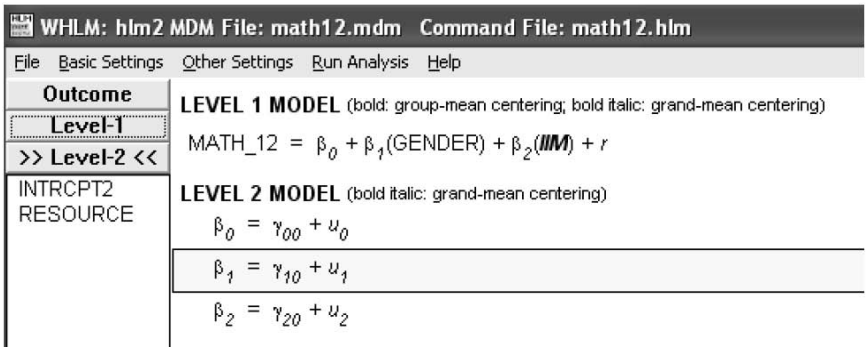


FIGURE 8.13 Adding a second level one predictor to a two-level model in HLM.

and the level two equation is:

$$\begin{cases} \beta_{0j} = \gamma_{00} + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \\ \beta_{2j} = \gamma_{20} + u_{2j} \end{cases} \quad (10)$$

This will appear in the HLM command window without i and j subscripts as can be seen in Figure 8.13 below.

The user can see that the centering used for *Gender* differs from that used for *IIM* from the different font style used in the HLM window for those variables in the model appearing in Figure 8.13 above. For uncentered variables, the variable name is not highlighted, for a group-mean centered variable, the variable appears in bold font and for a grand-mean centered variable, the variable's name is bolded and italicized.

The fixed effects results for the model contained in Equations 9 and 10 are as follows:

Final estimation of fixed effects (with robust standard errors)						
Fixed Effect		Coefficient	Standard Error	T-ratio	Approx. df	P-Value
For INTRCPT1, INTRCPT2, G00	B0	92.717182	0.756761	122.518	29	0.000
For GENDER slope, INTRCPT2, G10	B1	10.750966	0.900732	11.936	29	0.000
For IIM slope, INTRCPT2, G20	B2	0.552540	0.102726	5.379	29	0.000

Now the intercept represents the average *Math_12* score for a boy with an *IIM* score equal to that of the sample's mean *IIM* score. The intercept is significantly greater than zero ($\hat{\gamma}_{00} = 92.72$, $t(29) = 122.52$, $p < .0001$). There is a significant *Gender* effect favoring girls over boys ($\hat{\gamma}_{10} = 10.75$, $t(29) = 11.94$, $p < .0001$). The magnitude of this gender effect indicates that girls are predicted to have scores over 10 points higher on the *Math_12* than do boys with the same *IIM* score. The coefficient for *IIM* is also significantly greater than zero ($\hat{\gamma}_{20} = 5.38$, $t(29) = 5.38$, $p < .0001$) indicating a strong positive relationship between students' interest in mathematics and their performance on the *Math_12*.

The table of random effects' estimates from the HLM output appears below:

Final estimation of variance components:						
Random Effect		Standard Deviation	Variance Component	df	Chi- square	P- value
INTRCPT1,	U0	2.24432	5.03699	18	12.50906	>.500
GENDER slope,	U1	1.40596	1.97674	18	21.80177	0.240
IIM slope,	U2	0.29608	0.08766	18	29.18635	0.046
level-1,	R	4.21052	17.72844			

The estimate of the remaining level one variability is now 17.73 indicating that the addition of *Gender* has explained an additional 16.43% of the variability within classrooms (originally 50.47 in the unconditional model, down to 26.02 in the conditional model with *IIM* only as a predictor). Only 13.31% of the level one variability remains unexplained. The information contained in the table seems to indicate that there is not a significant amount of level two (among-classrooms) variability in the intercept or the *Gender* coefficient ($p > .05$). It should be emphasized that due to the small sample size within groups (i.e. the average number of children per classroom in this dataset is only 4.5) there is only low statistical power for estimation of the random effects (Hox, 2002).

The deviance is 794.63 with seven parameters estimated (three variances of random effects: u_{0j} , u_{1j} , u_{2j} , three covariances between the three random effects and σ^2). The difference in the deviances between this model and the one including only *IIM* is 85.47 which is still statistically significant ($\chi^2(3) = 85.47$, $p < .0001$) indicating that there would be a significant decrease in fit with *Gender* not included in the model.

Despite the lack of significance in the level two variability and due to the likely lack of statistical power in the dataset for identifying remaining level two variability (as well as for pedagogical purposes), addition of a level two (classroom) predictor to the model will now be demonstrated.

8.9 ADDITION OF A LEVEL TWO PREDICTOR TO A TWO-LEVEL HLM

In the classroom dataset, there was a measure of each of the classroom's mathematics pedagogy resources (*Resource*). It was hypothesized that there was a positive relationship between the amount of such resources in a classroom and the class's performance on the *Math_12* controlling for gender differences and for students' interest in mathematics. This translates into a hypothesis that *Resource* would predict some of the variability in the intercept. The original level one equa-

tion (see Equation 9) will remain unchanged. However, the set of level two equations (Equation 10) needs to be modified to include *Resource* as a predictor of β_{0j} , such that:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}Resource_j + u_{0j} \\ +\beta_{1j} = \gamma_{10} + u_{1j} \\ \beta_{2j} = \gamma_{20} + u_{2j} \end{cases} \quad (11)$$

To accomplish this in HLM, the user must first click on the relevant level two equation (the first of the three listed in Equation 11). In the current example, the user is interested in adding the level two variable to the intercept equation (the one for β_{0j}) so the user should make sure that that equation is highlighted. Then the user should click on the “Level 2” button in the upper left corner to call up the possible level two variables. Only one variable, *Resource*, can be added. (It should be noted here that the default in HLM is to include an intercept in the model. This default can be over-ridden by clicking on the relevant Intercept variable. See the HLM manual for further details). Once the user has clicked on *Resource*, the type of centering for the variable must be selected (from uncentered or grand-mean centered). Grand-mean centering will be selected so that the coefficient, γ_{01} , can be interpreted as describing a classroom with an average amount of resources. Once this is achieved, the HLM command screen appears as in Figure 8.14 below.

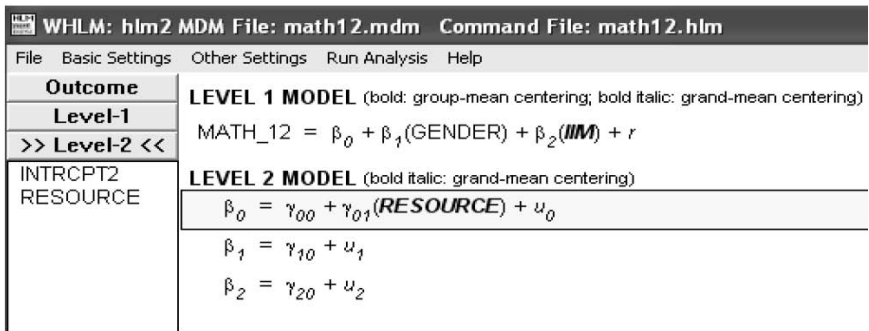


FIGURE 8.14 Adding a level two predictor to a two-level model in HLM.

Once the output file has been specified in “Basic Specifications” and the command file saved, the analysis can be run. More iterations are needed than are speci-

```

The value of the likelihood function at iteration 78 = -3.940063E+002
The value of the likelihood function at iteration 79 = -3.940061E+002
The value of the likelihood function at iteration 80 = -3.940059E+002
The value of the likelihood function at iteration 81 = -3.940058E+002
The value of the likelihood function at iteration 82 = -3.940056E+002
The value of the likelihood function at iteration 83 = -3.940054E+002
The value of the likelihood function at iteration 84 = -3.940053E+002
The value of the likelihood function at iteration 85 = -3.940051E+002
The value of the likelihood function at iteration 86 = -3.940049E+002
The value of the likelihood function at iteration 87 = -3.940048E+002
The value of the likelihood function at iteration 88 = -3.940046E+002
The value of the likelihood function at iteration 89 = -3.940045E+002
The value of the likelihood function at iteration 90 = -3.940043E+002
The value of the likelihood function at iteration 91 = -3.940041E+002
The value of the likelihood function at iteration 92 = -3.940040E+002
The value of the likelihood function at iteration 93 = -3.940038E+002
The value of the likelihood function at iteration 94 = -3.940037E+002
The value of the likelihood function at iteration 95 = -3.940035E+002
The value of the likelihood function at iteration 96 = -3.940034E+002
The value of the likelihood function at iteration 97 = -3.940032E+002
The value of the likelihood function at iteration 98 = -3.940031E+002
The value of the likelihood function at iteration 99 = -3.940029E+002

The maximum number of iterations has been reached, but the analysis has
not converged. Do you want to continue until convergence? _

```

FIGURE 8.15 MS-DOS window for a solution that was slow to converge.

fied as the default (100) as evidenced in the MS-DOS window that resulted (and is presented in Figure 8.15).

The user is prompted at the bottom of the screen that the program will continue its iterations towards estimation of a final solution if the user so desires. The user should enter “Y” if they are willing to wait through additional iterations. It should be noted that the solution can be considered more stable with fewer iterations. In addition, the estimation of multiple random effects with possibly insufficient sample size can aggravate the location of a solution. Should the user be prompted to use additional iterations, the user might wish to continue with the solution but change the model to re-estimate it by modeling one or several of the parameters as fixed instead of random.

When the model’s estimation did converge after 1497 iterations, the additional fixed effect estimate, γ_{01} , appears in the output on the following page:

Final estimation of fixed effects
(with robust standard errors)

Fixed Effect		Coefficient	Standard Error	T-ratio	Approx. df	P-Value
For INTRCPT1,	B0					
INTRCPT2, G00		92.716978	0.632195	146.659	28	0.000
RESOURCE, G01		1.416373	0.605262	2.340	28	0.027
For GENDER slope,	B1					
INTRCPT2, G10		10.612002	0.852843	12.443	29	0.000
For IIM slope,	B2					
INTRCPT2, G20		0.598363	0.097142	6.160	29	0.000

The classroom *Resource* measure was found to be significantly positively related to *Math_12* controlling for gender and *IIM* ($\hat{\gamma}_{01} = 1.42$, $t(28) = 2.34$, $p < .05$). From the random effects' estimates output:

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1,	U0	1.54086	2.37425	17	11.53940	>.500
GENDER slope,	U1	0.78565	0.61724	18	21.40071	0.259
IIM slope,	U2	0.24638	0.06070	18	27.74502	0.066
level-1,	R	4.22362	17.83898			

The addition of *Resource* has reduced the level two variability in the intercept from 5.04 (in the model that included *Gender* and *IIM*) to 2.37. The deviance of the current model in which seven covariance parameters were estimated was 787.94.

8.10 EVALUATING THE EFFICACY OF A TREATMENT

HLM can be used to evaluate whether two or more counseling (or, say teaching) methods have a differential effect on some outcome. This example is designed to investigate the impact of two counseling methods and whether they have a differential effect on empathy. It should be noted that in this example a smaller sample size is used than is typically recommended for HLM analyses. This is done to facilitate its presentation. Five groups of patients are treated with each counseling method. Each group has four patients. While groups are nested within counseling method, because the research question is about a comparison of the two counseling

Level 1				Level 2	
PatId	Gp	Emp	Content	Gp	Couns
1	1	23	30	1	0
2	1	22	33	2	0
3	1	20	30	3	0
4	1	19	28	4	0
5	2	16	19	5	0
6	2	17	21	6	1
7	2	18	28	7	1
8	2	19	37	8	1
9	3	25	35	9	1
10	3	28	38	10	1
11	3	29	38		
12	3	31	37		
13	4	27	44		
14	4	23	30		
15	4	22	31		
16	4	21	25		
17	5	32	37		
18	5	31	46		
19	5	28	42		
20	5	26	39		
21	6	13	24		
22	6	12	19		
23	6	14	31		
24	6	15	25		
25	7	16	27		
26	7	17	34		
27	7	14	24		
28	7	12	22		
29	8	11	25		
30	8	10	17		
31	8	20	31		
32	8	15	30		
33	9	21	26		
34	9	18	28		
35	9	19	27		
36	9	23	33		
37	10	18	24		
38	10	17	33		
39	10	16	33		
40	10	23	29		

methods, they do not constitute a clustering level. Thus, we have a two-level nested design, with patients (level one) nested within groups (level two) and counseling method used as a fixed level two (group-level) variable. Counseling method will be used as a predictor to explain some of the variability between groups. **We have two**

separate data files with group ID (*gp*) in both files. The level one file contains group ID with data (scores on the empathy scale and the patient's id number) for the four subjects in each of the ten groups. In addition, the level one file includes a measure of the patient's contentment (*content*). The level two file has the group ID variable along with the counseling method (*couns*) employed in the relevant group coded either as 0 or 1. The data files are presented on p. 352.

The MDM file is constructed and then the analysis conducted using HLM6. The model estimated includes counseling method as a fixed predictor. No level one predictors are included in the model. The HLM results are presented below:

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. df	P-value
For INTRCPT1, B0					
INTRCPT2, G00	23.850000	1.825171	13.067	8	0.000
COUNS, G01	-7.650000	2.581182	-2.964	8	0.019

The outcome variable is EMP

Final estimation of fixed effects
(with robust standard errors)

Fixed Effect	Coefficient	Standard Error	T-ratio	Approx. df	P-value
For INTRCPT1, B0					
INTRCPT2, G00	23.850000	1.973069	12.088	8	0.000
COUNS, G01	-7.650000	2.308679	-3.314	8	0.012

The robust standard errors are appropriate for datasets having a moderate to large number of level 2 units. These data do not meet this criterion.

Final estimation of variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, U0	3.86868	14.96667	8	78.86560	0.000
level-1, R	2.59968	6.75833			

Statistics for current covariance components model

Deviance = 204.746721

Number of estimated parameters = 2

As noted in the output, there is an insufficient number of level two (group) units and thus the results with robust standard errors should not be used here. Note that the counseling method effect results [$t(8) = -2.964, p = .019$] indicate that the counseling method is statistically significant with the method coded using a zero having a stronger impact on empathy than the method coded with a one. Note also that the degrees of freedom is 8 which corresponds to the degrees of freedom between groups for a regular ANOVA. In their text, Maxwell and Delaney (p. 514, 2004) note that the proper error term for a nested design such as this is groups within methods. This is what would be used if SPSS had been used to analyze the data. Control lines and selected output from an SPSS analysis is given below:

SPSS Control Lines for Univariate Nested Design

```
DATA LIST FREE/COUNS GP SUB EMP.
BEGIN DATA.
0 1 1 23    0 1 2 22    0 1 3 20    0 1 4 19
0 2 1 16    0 2 2 17    0 2 3 18    0 2 4 19
0 3 1 25    0 3 2 28    0 3 3 29    0 3 4 31
0 4 1 27    0 4 2 23    0 4 3 22    0 4 4 21
0 5 1 32    0 5 2 31    0 5 3 28    0 5 4 26
1 6 1 13    1 6 2 12    1 6 3 14    1 6 4 15
1 7 1 16    1 7 2 17    1 7 3 14    1 7 4 12
1 8 1 11    1 8 2 10    1 8 3 20    1 8 4 15
1 9 1 21    1 9 2 18    1 9 3 19    1 9 4 23
1 10 1 18   1 10 2 17    1 10 3 16    1 10 4 23
END DATA.
UNIANOVA EMP BY COUNS GP SUB/
RANDOM GP SUB/
DESIGN SUB(GP(COUNS)) GP(COUNS) COUNS /
PRINT = DESCRIPTIVES/.
```

Note that in the SPSS syntax, the first number indicates the counseling method (0 or 1), the second number the group the patient is in (1 through 10), and the third number indicates the subject number. Thus the first set of four numbers represents that the subject is in the counseling method coded with a 0, is the first person in group 1 and has an empathy score of 23. The “RANDOM GP SUB/” line indicates that group and subject are being modeled as random factors. Lastly, the DESIGN command line indicates a nested design, with patients nested within groups which in turn are nested within counseling methods.

SPSS Printout for Three-Level Nested Design

Tests of Between-Subjects Effects						
Dependent Variable: EMP						
Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	16040.025	1	16040.025	240.751	.000
	Error	533.000	8	66.625(a)		
SUB(GP(COUN S))	Hypothesis	202.750	30	6.758	9.858	.000
	Error	.000	0	.(b)		
GP(COUNS)	Hypothesis	533.000	8	66.625	8.784	.018
	Error	202.750	30	6.758(c)		
COUNS	Hypothesis	585.225	1	585.225	8.784	.018
	Error	533.000	8	66.625(a)		

a MS(GP(COUNS))
b MS(Error)
c MS(SUB(GP(COUNS)))

Note that the error term for the counseling method effect is groups within methods. Remember that there are 5 groups within each of the two counseling methods so that the degrees of freedom is 8. This can be seen on the SPSS output above where $F = 8.784$, $p = .018$ for the counseling effect (*couns*). This corresponds (with rounding error) to the square of the effect found with HLM for the *couns* variable: $(-2.964)^2 = 8.785$ indicating the correspondence between the SPSS and HLM analysis for this fixed effect. However, the error term for groups within counseling in SPSS is NOT correct because it is based on 30 degrees of freedom (for the error term). The degrees of freedom for error SHOULD be less than 30 because the observations within the groups are dependent. Here, one would prefer the results from the HLM6 analysis, which indicates significant group variability ($\chi^2 = 78.866$, $p < .05$). Note lastly that for an analysis of three counseling methods, two dummy variables would be needed to identify group membership and both of these variables could be used as predictors at level 2.

Adding a Level One Predictor to the Empathy Model Data

In the model estimated for the *Empathy* data above where level one is formulated:

$$Emp_{ij} = \beta_{0j} + r_{ij}$$

and level two:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Tx_j + u_{0j},$$

the variability in the intercept across treatment groups (τ_{00}) even after controlling for treatment effects is seen to be significantly greater than zero [$\chi^2(8) = 78.86560$,

$p < .05$]. A researcher might be interested in adding a level one predictor to help explain some of this remaining variability in *Empathy* using the patient’s level of *Contentment* with the level one formulation becoming:

$$Emp_{ij} = \beta_{0j} + \beta_{1j}Content_{ij} + r_{ij}$$

and at level two:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}Tx_j + u_{0j} \\ \beta_{1j} = \gamma_{10} + u_{1j} \end{cases}$$

Addition of the level one predictor (*Content*) modifies interpretation of the intercept, γ_{00} , from the “predicted empathy score for patients in the group for which $Tx = 0$ ” to the “predicted empathy score for patients controlling for level of contentment (i.e. for whom *Content* = 0) in a treatment group for which $Tx = 0$. Note that we will grand-mean center *Content* so that a patient with *Content* = 0 is one at the mean on the contentment scale.

Estimating this model with HLM, we find the following fixed effect estimates:

Final estimation of fixed effects:						
Fixed Effect		Coefficient	Standard Error	T-ratio	Approx. df	P-value
For INTRCPT1, INTRCPT2, TX, G01	B0					
	G00	22.584852	1.228768	18.380	8	0.000
		-5.318897	1.719344	-3.094	8	0.016
For CON slope, INTRCPT2, G10	B1					
	G10	0.355810	0.073244	4.858	9	0.001

We see that the coefficient for *Content* is statistically significant ($\gamma_{10} = 0.356$, $t(9) = 4.86$, $p < .05$). We can also see that a treatment effect is still found to favor the groups for whom $Tx = 0$ ($\gamma_{01} = -5.319$, $t(8) = -3.094$, $p < .05$).

The random effects estimates were:

Final estimation of variance components:						
Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
INTRCPT1, U0		2.44137	5.96028	8	29.46502	0.000
CON slope, U1		0.04038	0.00163	9	9.39983	0.401
level-1, R		2.22379	4.94526			

For this model in which *Content* is modeled to vary randomly across therapy groups we can thus see that a significant amount of variability remains in the intercept (even with *Content* added to the model). However, there is not a significant amount of variability between therapy groups in the relationship between patients' *Content* and their *Empathy* scores. Thus our final model will include *Content* modeled as an effect that is fixed across therapy groups such that level two, we model:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}Tx_j + u_{0j} \\ \beta_{1j} = \gamma_{10} \end{cases}$$

The fixed effects estimates were:

Final estimation of fixed effects:							
Fixed Effect			Coefficient	Standard Error	T-ratio	Approx. df	P-value
For	INTRCPT1,	B0					
	INTRCPT2, G00		22.773094	1.249052	18.232	8	0.000
	TX, G01		-5.496188	1.796275	-3.060	8	0.017
For	CON slope,	B1					
	INTRCPT2, G10		0.341875	0.073204	4.670	37	0.000

These parameter estimates can be substituted into the level two formulation:

$$\begin{cases} \beta_{0j} = 22.77 - 5.50Tx_j + u_{0j} \\ \beta_{1j} = 0.34 \end{cases}$$

To facilitate interpretation of the results, it can help to obtain the single equation (by substituting the level two equations for β_{0j} and β_{1j} into the level one equation to obtain:

$$Emp_{ij} = 22.77 - 5.50Tx_j + 0.34Content_{ij} + r_{ij} + u_{0j}$$

and then (as can be done with simple regression), values for the predictors can be substituted into this single equation. For example, substituting the relevant values into the single multilevel equation, the combinations of *Tx* and *Content* scores results in the predicted *Empathy* scores that appear in the following table:

	$Tx = 0$	$Tx = 1$
$Content = 0$	22.77	17.27
$Content = 1$	23.11	17.61
$Content = 2$	23.45	17.95

Thus, for example, the value for γ_{00} represents the predicted *Emp* score when $Tx = 0$ and for someone at the mean on *Content* (i.e. for someone for whom $Content = 0$). The Tx coefficient represents the treatment's effect on *Emp* controlling for *Content* levels. In other words, for two participants with the same *Content* score, one of whom is in the $Tx = 0$ group while the other is in the $Tx = 1$ group, there will be a predicted difference of 5.5 points on the *Emp* scale (with the difference favoring the $Tx = 0$ member). In the table above, the difference for two people with $Content = 0$ is $22.77 - 17.27 = 5.5$. Similarly for two people with $Content = 2$ (2 points above the mean on *Content*), the predicted difference for those in $Tx = 0$ versus $Tx = 1$ is $23.45 - 17.95$. Lastly, the *Content* coefficient indicates that for two patients in the same Tx group, a difference of one on the *Content* scale is associated with an *Emp* score predicted to be .34 points higher. In other words, controlling for the treatment effect, the more contented a patient, the better their *Empathy* is anticipated to be. Thus we see in the table that for two people in the $Tx = 0$ group, one with $Content = 1$ and the other patient with $Content = 0$, the difference in their predicted *Emp* scores is: $23.11 - 22.77 = 0.34$. Similarly for two people in $Tx = 1$, one with $Content = 2$, the other with $Content = 1$, the predicted difference in *Emp* is: $17.95 - 17.61$

8.11 FINAL COMMENTS ON HLM

It should be emphasized that this chapter has provided only a very brief introduction to multilevel modeling and the use of HLM software to estimate the model parameters. It should be also be noted that despite the ease with which researchers can use software such as HLM to estimate their multilevel models, it behooves the user to ensure that they understand the model being estimated, how to interpret the resulting parameter estimates and associated significance tests as well as the appropriateness of the assumptions made. While not demonstrated in this chapter, because this is an introductory treatment, a residual file can be easily created. As Raudenbush et al. note on p. 13 of the HLM5 manual and repeat on p. 15 of the HLM6 manual:

"After fitting a hierarchical model, it is wise to check the tenability of the assumptions underlying the model:

Are the distributional assumptions realistic?
 Are results likely to be affected by outliers or influential observations?
 Have important variables been omitted or nonlinear relationships been ignored?"

HLM software can be used to provide the residuals for models estimated.

Aside from HLM, there are several other software programs that can be used to estimate multilevel models including MLwiN (Goldstein, et al., 1998), SAS Proc Mixed (Littell, Milliken, Stroup, & Wolfinger, 1996; see Singer (1998) for a well-written introductory article describing use of PROC MIXED for multilevel modeling) and VARCL (Longford, 1988) among others. Even the latest versions of SPSS include some basic hierarchical modeling estimation routines. Kreft and De Leeuw (1998) provide some good descriptions of the available multilevel programs as well as website references for the interested user.

The list of multilevel textbooks provided earlier in the chapter can provide the reader with more detailed worked examples as well as fuller descriptions of the estimation used and the assumptions made when analyzing these multilevel models. In addition, the texts provide excellent resources for the reader to find out about more advanced multilevel modeling techniques including models with dichotomous or ordinal outcomes, models with multivariate outcomes, meta-analytic models, and models for use with cross-classified data structures.

The same caveats that apply to model-building using simple single-level regression analyses apply to model-building with multilevel models. Choosing a final model based on resulting estimates from a series of models can lead to selection of a model that is very sample-specific. As with any kind of model-fitting, if the analyst has a large enough sample, then the data can be randomly divided to provide a cross-validation sample to use to test the final model selected based on results from the other half of the sample (Hox, 2002).

It is hoped that researchers continue to become more familiar with the complexities of multilevel modeling and that they will be increasingly applied for the analysis of relevant data structures.

The data files for the MATH_12 example follow. There are 135 students and 30 teachers in this artificial data set.

2lvl_student_L1

	<i>teachid</i>	<i>studid</i>	<i>gender</i>	<i>math_12</i>	<i>iim</i>
1	1	5	0	94	35
2	1	7	1	107	35
3	1	9	0	97	42
4	2	14	0	92	42
5	2	16	0	94	39
6	2	19	0	92	49
7	2	20	1	105	41
8	3	21	0	93	40
9	3	28	1	101	50
10	3	29	1	113	45
11	4	32	1	103	42
12	4	35	1	101	43
13	4	36	1	108	47
14	5	47	0	98	37
15	5	49	0	95	37
16	5	50	1	99	32
17	6	52	0	78	33
18	6	58	0	83	38
19	6	60	0	81	34
20	7	61	0	96	40
21	7	68	0	93	43
22	7	69	1	99	37
23	8	72	0	92	38
24	8	74	0	93	43
25	8	76	0	77	36
26	8	78	0	98	43
27	8	80	1	99	47
28	9	81	0	95	49
29	9	82	0	96	47
30	9	83	0	98	48
31	9	84	0	93	45
32	9	90	0	94	49
33	10	92	1	105	46
34	10	94	1	104	45
35	10	96	1	104	45
36	10	97	1	107	46
37	10	100	1	110	47
38	11	101	1	107	39
39	11	105	1	97	36
40	11	107	1	101	40
41	11	108	0	82	36
42	11	109	0	86	34
43	12	112	1	100	41
44	12	115	1	105	51
45	12	117	0	94	44

(Continued)

	<i>teachid</i>	<i>studid</i>	<i>gender</i>	<i>math_12</i>	<i>iim</i>
46	12	118	1	116	53
47	12	119	1	108	49
48	12	120	1	99	51
49	13	123	1	106	41
50	13	125	1	116	50
51	13	128	0	96	40
52	13	129	1	105	44
53	13	130	1	96	37
54	14	132	1	105	44
55	14	133	1	105	42
56	14	139	1	111	41
57	14	140	1	112	48
58	15	141	1	99	41
59	15	142	0	93	40
60	15	144	1	106	42
61	15	145	1	106	44
62	15	149	1	108	44
63	16	151	1	106	48
64	16	154	1	106	43
65	16	155	1	99	45
66	16	157	0	91	41
67	16	158	1	98	45
68	17	161	0	91	32
69	17	164	0	83	30
70	17	165	1	117	48
71	17	166	1	102	43
72	17	169	1	116	53
73	17	170	1	105	46
74	18	171	0	96	42
75	18	172	1	104	45
76	18	174	0	78	33
77	18	175	0	94	43
78	18	176	0	95	37
79	18	177	0	81	33
80	18	178	0	88	36
81	18	179	0	96	37
82	18	180	0	95	35
83	19	181	0	96	45
84	19	183	0	91	45
85	19	185	0	93	48
86	19	187	0	87	44
87	19	188	0	83	51
88	19	190	0	83	43
89	20	194	1	101	43
90	20	199	0	85	37

(Continued)

	<i>teachid</i>	<i>studid</i>	<i>gender</i>	<i>math_12</i>	<i>iim</i>
91	20	200	0	92	38
92	21	205	1	101	45
93	21	206	1	104	40
94	21	207	1	100	44
95	21	208	1	101	47
96	21	210	1	102	41
97	22	212	1	113	44
98	22	214	1	100	39
99	22	216	0	97	49
100	22	218	1	109	35
101	22	219	1	111	42
102	22	220	1	99	37
103	23	221	0	92	42
104	23	223	0	90	33
105	23	224	0	95	37
106	23	225	1	105	41
107	24	233	0	98	38
108	24	234	1	103	43
109	24	235	1	101	42
110	24	236	0	96	37
111	24	238	1	102	40
112	25	243	0	90	42
113	25	246	0	90	45
114	25	247	0	88	41
115	25	249	0	95	42
116	26	254	0	96	45
117	26	256	1	101	43
118	26	258	1	107	44
119	26	259	1	104	40
120	26	260	1	99	42
121	27	261	1	108	46
122	27	263	0	85	38
123	27	266	0	91	39
124	27	267	0	89	38
125	27	270	1	97	38
126	28	271	1	108	56
127	28	272	1	97	41
128	28	273	1	111	49
129	28	274	1	105	50
130	28	280	1	109	45
131	29	287	0	93	40
132	29	290	0	82	30
133	30	297	1	108	45
134	30	298	1	104	38
135	30	299	0	84	24

2lvl_class_L2

	<i>teachid</i>	<i>resource</i>
1	1	7
2	2	7
3	3	5
4	4	7
5	5	7
6	6	5
7	7	5
8	8	5
9	9	5
10	10	6
11	11	5
12	12	6
13	13	5
14	14	6
15	15	5
16	16	5
17	17	6
18	18	7
19	19	4
20	20	6
21	21	6
22	22	5
23	23	5
24	24	6
25	25	5
26	26	6
27	27	6
28	28	5
29	29	6
30	30	7

Appendix A

Data Sets

CONTENTS

- A.1 Clinical Data
- A.2 Alcoholics Data
- A.3 Sesame Street Data
- A.4 Headache Data
- A.5 Cartoon Data
- A.6 Attitude Data
- A.7 National Academy of Sciences Data
- A.8 Agresti Home Sales Data

CLINICAL DATA

The data for this study was drawn from the archives of Children's Hospital Medical Center's Department of Psychology in Cincinnati, Ohio. Thirty seven subjects were eventually selected from each of three diagnostic groups:

1. Encopretic children: These children have problems with fecal soiling. Clinically, parents report that the child "forgets," is too engrossed in other activities, or delays in going to the toilet. Problems that have been found to be associated with encopresis include anxiousness, social withdrawal, motor integration, and attention deficit.
2. Hyperactive children.
3. General clinic group: adjustment disorder, disturbed.

The selection criteria for the subjects were: male, 7–14 years of age, Full Scale WISC-R score above 85, and an absence of any concurrent disability or condition that could account for bowel control or attention problems.

Factor scores on the WISC-R were used to assess cognitive development and attention/information intake processes. The factor scores are for verbal comprehen-

sion (VERBCOMP), perceptual organization (PERCORG), and freedom from distractability (FREEDIST).

CLINICAL DATA							
G P I D	V	P	F	G P I D	V	P	F
	E	R	R		E	R	R
	R	E	E		R	E	E
	B	R	E		B	R	E
	C	C	D		C	C	D
P	O	O	I	P	O	O	I
I	M	R	S	I	M	R	S
D	P	G	T	D	P	G	T
1	7.25	11.00	9.00	2	7.75	11.50	8.33
1	13.50	13.75	9.67	2	9.00	8.75	8.67
1	8.00	8.25	6.67	2	7.50	9.00	7.33
1	8.50	11.75	9.00	2	8.00	9.75	5.33
1	6.25	12.75	7.67	2	14.00	10.25	11.00
1	12.50	14.00	10.33	2	10.00	11.50	8.00
1	11.00	13.25	8.67	2	11.75	6.50	8.33
1	11.25	10.75	9.67	2	9.75	9.25	8.67
1	11.75	13.75	14.67	2	10.00	8.75	7.67
1	7.50	12.50	6.33	2	10.50	13.50	12.00
1	8.00	11.00	7.67	2	9.25	8.25	8.33
1	10.25	11.50	7.00	2	9.75	10.25	6.00
1	9.75	13.25	8.00	2	11.50	9.00	11.33
1	9.00	11.00	7.33	2	12.50	12.50	9.67
1	10.00	10.50	10.00	2	13.75	12.50	11.00
1	12.00	15.25	12.67	2	11.00	11.25	9.33
1	9.50	8.50	6.00	2	11.00	8.25	10.00
1	13.25	10.75	8.67	2	9.00	11.25	9.67
1	10.50	11.50	5.33	3	8.50	8.00	9.00
1	10.50	9.25	8.67	3	9.50	10.50	6.67
1	10.50	11.75	7.67	3	8.00	10.25	9.67
1	8.25	11.00	5.67	3	9.25	11.50	5.33
1	7.75	9.25	8.00	3	9.25	9.75	5.33
1	12.75	13.75	12.00	3	15.50	12.50	14.00
1	8.75	9.75	9.00	3	9.25	11.00	10.00
1	7.75	8.75	5.00	3	8.25	10.50	7.00
1	9.50	12.00	7.00	3	11.50	13.00	8.00
1	10.50	11.50	5.33	3	9.25	10.75	8.67
1	15.50	14.00	8.00	3	10.00	10.50	9.33
1	9.75	9.75	6.33	3	12.50	12.00	8.33
1	9.50	12.50	9.33	3	10.00	13.25	9.67
1	8.00	7.50	5.00	3	10.75	12.50	9.00
1	14.75	12.25	8.33	3	11.25	13.50	8.00
1	9.25	9.50	9.00	3	12.25	10.75	9.00

(Continued)

CLINICAL DATA (Continued)

G P I D	V	P	F	G P I D	V	P	F
	E	R	R		E	R	R
	R	E	E		R	E	E
	B	R	E		B	R	E
	C	C	D		C	C	D
O	O	I		O	O	I	
M	R	S		M	R	S	
P	G	T		P	G	T	
1	9.25	11.75	7.33	3	16.00	13.25	12.67
1	8.50	9.50	8.00	3	12.25	13.75	14.00
1	12.25	10.75	9.00	3	11.00	11.00	9.67
2	9.75	10.25	11.00	3	12.25	12.00	10.33
2	9.50	11.75	11.33	3	6.50	8.25	10.33
2	10.75	12.00	9.67	3	8.75	9.75	10.33
2	9.00	11.50	7.67	3	14.25	16.00	16.33
2	11.25	13.75	12.33	3	6.00	10.00	7.67
2	12.25	11.75	12.00	3	14.00	13.50	9.00
2	12.00	8.75	5.67	3	12.00	9.25	8.00
2	11.00	10.25	7.00	3	10.00	9.50	10.33
2	10.25	8.75	6.33	3	12.50	12.75	9.67
2	12.50	13.75	9.00	3	11.50	12.75	10.33
2	10.25	13.25	8.00	3	10.75	10.75	8.33
2	12.75	13.25	11.67	3	9.00	10.75	11.33
2	6.75	9.50	7.33	3	11.50	13.00	8.00
2	9.00	10.00	8.00	3	11.25	14.25	11.33
2	9.25	9.75	10.00	3	10.00	10.50	9.33
2	13.75	13.25	7.33	3	11.25	12.50	13.00
2	9.00	8.50	11.33	3	8.00	10.25	9.67
2	9.75	7.25	8.67	3	9.75	10.50	9.67
2	8.25	10.25	7.33				

DESCRIPTION OF ALCOHOLICS DATA SET

This is data from a study into the causes of relapse among alcoholics conducted at an Addiction Research Unit in London, England. The 251 subjects are those who presented themselves for treatment of alcoholism at several hospitals and related agencies. The subjects were divided into three groups;

- Group 1: those never having previously experienced relapse after trying to give up heavy drinking.
- Group 2: those who claimed to have relapsed, but no more than two or three times.
- Group 3: those who had a longer history of relapse of four or more times.

The dependent variables for the study came from a Relapse Inventory Precipitants Inventory, which the subjects were asked to fill out at the point of admission into treatment. This inventory was developed from work on a previous survey that had identified three areas of vulnerability to relapse, the measurement of which yielded the following dependent variables:

UNPLMOOD—unpleasant mood states; for example, depression.

EUPHORIC—euphoric states and related situations; for example, celebrations and parties.

LESSVIGL—an area designated as lessened vigilance, for example, a temptation to believe that one or more drinks would cause no problem.

The grouping variable is given first for each subject, with UNPLMOOD, EUPHORIC, and LESSVIGL in that order.

ALCOHOLICS DATA

<i>Grp</i>	<i>UM</i>	<i>ES</i>	<i>LV</i>	<i>Grp</i>	<i>UM</i>	<i>ES</i>	<i>LV</i>
2	27	0	3	2	12	14	4
1	37	19	6	1	31	13	9
2	2	2	6	2	2	3	3
3	24	15	3	3	3	4	0
1	34	13	4	1	26	9	9
1	37	10	7	1	33	14	9
1	34	23	9	1	18	7	6
3	7	9	7	1	16	8	9
3	10	0	0	2	13	9	7
3	6	7	6	1	15	12	3
2	30	16	9	1	25	18	6
3	24	21	8	1	23	11	4
1	22	6	8	1	26	6	9
1	24	7	7	1	18	0	2
1	21	18	6	2	6	6	4
1	0	4	3	1	19	13	6
1	14	7	2	1	34	23	9
2	14	13	0	1	26	21	7
1	10	13	7	2	28	22	7
1	13	18	9	2	36	21	9
3	41	16	9	1	19	16	9
1	24	15	3	1	32	18	8
2	1	6	3	2	3	5	4
2	15	16	7	1	9	16	9
2	6	2	0	2	0	6	7
1	34	16	6	1	8	4	5
1	37	19	9	1	42	22	9
2	2	4	0	1	41	21	8
1	28	6	9	2	16	10	2
2	21	15	8	2	7	12	4
3	13	3	3	2	36	11	7
1	32	9	8	3	18	10	9
2	25	2	9	1	24	9	8
3	0	0	0	1	37	19	9
2	27	13	4	2	3	2	0
1	12	10	4	1	16	12	8
2	11	11	7	2	14	5	7
1	30	13	6	1	35	16	6
2	32	19	5	1	35	15	7
1	9	7	1	1	34	21	9
1	32	24	9	1	34	23	7
2	31	18	5	2	31	7	9
3	26	16	3	1	35	10	5
3	0	2	1	1	24	7	6
1	5	3	4	1	29	16	9

(Continued)

ALCOHOLICS DATA

<i>Grp</i>	<i>UM</i>	<i>ES</i>	<i>LV</i>	<i>Grp</i>	<i>UM</i>	<i>ES</i>	<i>LV</i>
1	15	5	3	1	17	3	9
1	29	18	4	1	11	12	6
1	28	21	9	2	39	23	9
2	16	14	9	1	21	10	9
3	15	5	8	1	11	3	4
1	27	15	7	2	33	17	9
2	21	12	5	2	0	0	0
1	42	23	9	2	31	14	7
2	21	12	9	1	18	14	2
1	16	0	3	3	41	12	9
1	26	0	9	1	5	10	9
1	20	2	5	3	30	12	6
1	1	1	1	2	3	9	6
3	15	23	9	2	9	8	2
1	16	13	3	1	17	20	7
3	0	10	3	2	15	3	6
2	22	10	5	1	28	21	6
3	20	6	1	1	14	10	2
3	14	3	4	2	13	14	7
1	7	7	3	2	0	0	0
3	14	12	4	2	0	6	2
2	14	1	4	1	6	1	2
2	40	5	4	1	35	16	9
1	24	9	7	1	23	3	8
2	15	5	5	3	10	5	6
1	8	17	6	1	29	20	9
2	5	4	3	1	23	12	6
1	25	1	9	2	4	0	2
2	17	2	9	1	11	13	8
1	17	18	8	2	34	16	7
1	26	18	8	3	0	0	0
1	29	10	6	1	26	5	8
1	26	15	9	3	15	15	5
				1	3	5	2
2	20	14	9	3	34	11	7
1	10	17	9	2	6	19	6
3	31	20	6	1	18	12	5
1	18	7	3	1	3	5	7
1	22	13	4	1	35	15	9
2	21	9	6	2	17	5	5
2	3	7	4	2	29	16	7
2	16	1	0	2	32	20	8
1	15	12	7	1	13	5	3
3	7	5	1	2	37	22	9
2	18	12	4	1	26	13	6

(Continued)

ALCOHOLICS DATA (Continued)

<i>Grp</i>	<i>UM</i>	<i>ES</i>	<i>LV</i>	<i>Grp</i>	<i>UM</i>	<i>ES</i>	<i>LV</i>
1	27	11	9	1	17	9	4
1	18	9	8	1	28	10	5
2	0	0	0	3	28	8	6
1	8	14	6	3	5	3	0
2	7	7	6	2	25	19	8
1	7	4	5	3	19	23	3
2	19	21	8	2	33	14	5
2	30	16	8	2	7	17	4
3	6	16	8	1	19	13	6
1	28	13	9	1	32	13	9
1	17	8	3	3	0	3	0
2	17	13	6				
1	4	4	3	2	16	18	2
2	12	16	6	1	39	18	9
1	2	1	3	1	28	20	3
2	40	4	3	1	24	16	8
1	18	9	8	2	24	14	6
2	11	11	6	2	19	5	7
1	11	21	9	3	33	19	3
				1	15	10	7
2	23	12	7	1	24	16	7
1	32	20	9	2	33	14	7
2	14	3	3	1	26	3	5
2	10	6	4	2	26	12	8
3	32	7	4	1	26	16	6
				2	40	20	9
2	22	19	5	3	33	17	7
1	0	0	0	2	30	13	8
2	29	16	6	2	22	20	7
1	0	10	8	1	21	15	6
1	23	9	7	2	17	9	6
1	37	22	9	1	28	23	9
2	25	17	6	1	15	17	9
1	17	10	3	3	11	5	6
2	18	3	2	3	39	15	6
2	19	22	8	2	30	17	9
1	16	4	5	2	22	11	3
2	0	2	2				

DESCRIPTION OF SESAME STREET DATA BASE

This data is part of a large data set that evaluated the impact of the first year of the Sesame Street television series. Sesame Street was concerned mainly with teaching preschool related skills to children in the 3–5 year age range, with special emphasis on reaching 4 year old disadvantaged children. The format of the show was designed to hold young children's attention through action oriented, short duration presentations teaching specific preschool cognitive skills and some social skills. Each show was one hour and involved much repetition of concepts within and across shows.

A main concern for the evaluation, which was carried out at Educational Testing Service, was that it would permit generalization to the populations of children of most interest to the producers of the program (the Children's Television Workshop). Five populations were of interest:

1. Three to five year old disadvantaged children from inner city areas in various parts of the country.
2. Four year old advantaged suburban children.
3. Advantaged rural children.
4. Disadvantaged rural children.
5. Disadvantaged Spanish speaking children.

Children representative of these populations were sampled from five different sites in the United States.

Both before and after viewing the series the children were tested on a variety of cognitive variables (variables 8 through 19 in the data set), including knowledge of body parts, knowledge about letters, knowledge about numbers, etc.

The variables are arranged on the file as follows:

<i>Variable No.</i>	<i>Variable Name</i>	<i>Description</i>
1	ID	Subject identification number
2	SITE	Five different sampling sites coded as 1,2,3,4 or 5.
3	SEX	Male–1, Female–2
4	AGE	in months
5	VIEWCAT	Viewing categories coded as a 1 if children rarely watched the show to a 4 if the children watched the show on average of more than 5 times a week
6	SETTING	Setting in which Sesame Street was viewed, coded as 1 for home and coded as 2 for school
7	VIEWENC	A treatment condition in which some children were encouraged to view Sesame St (code–1) and others were not (code–2)
8	PREBODY	pretest on knowledge about body parts (maximum score–32)—naming and functions of body parts
9	PRELET	pretest on knowledge about letters (maximum score–58)—including recognizing letters, naming capital letters, matching letters in words
10	PREFORM	pretest on knowledge about forms (maximum score–20)—recognizing and naming forms
11	PRENUMB	pretest on knowledge about numbers (maximum score–54)—recognizing and naming numbers, counting, addition and subtraction
12	PRERELAT	pretest on knowledge of relational terms (maximum score–17)—amount, size and position relationship
13	PRECLASF	pretest on knowledge of classification skills (maximum score–24)—classifying by size, form, number and function
14	POSTBODY	posttest knowledge on body parts
15	POSTLET	posttest knowledge of letters
16	POSTFORM	posttest knowledge of forms
17	POSTNUMB	posttest knowledge of numbers
18	POSTREL	posttest knowledge of relations
19	POSTCLAS	posttest knowledge of classification skills
20	PEABODY	Mental age scores obtained from administration of the Peabody Picture Vocabulary Test as a pretest measure of vocabulary maturity

SESAME STREET DATA

				V I E T W E N N D Y				P R E F O U L A S D E R M B				P R E C L O T S F N T C L S				P O S T U R E A D			
I	S	A	C	I	E	O	E	R	E	L	A	S	D	E	R	M	E	A	D
D	E	X	E	T	G	C	T	M	B	T	F	Y	T	M	B	L	S	Y	
1	1	1	66	1	2	1	16	23	12	40	14	20	18	30	14	44	14	23	62
2	1	2	67	3	2	1	30	26	9	39	16	22	30	37	17	39	14	22	80
3	1	1	56	3	2	2	22	14	9	9	9	8	21	46	15	40	9	19	32
4	1	1	49	1	2	2	23	11	10	14	9	13	21	14	13	19	8	15	27
5	1	1	69	4	2	2	32	47	15	51	17	22	32	53	18	54	14	21	71
6	1	2	54	3	2	2	29	26	10	33	14	14	27	36	14	39	16	24	32
7	1	2	47	3	2	2	23	12	11	13	11	12	22	45	12	44	12	15	28
8	1	1	51	2	2	1	32	48	19	52	15	23	31	47	18	51	17	23	38
9	1	1	69	4	2	1	27	44	18	42	15	20	32	50	17	48	14	24	49
10	1	2	53	3	2	1	30	38	17	31	10	17	32	52	19	52	17	24	32
11	1	2	58	2	2	2	25	48	14	38	16	18	26	52	15	42	10	17	43
12	1	2	58	4	2	2	21	25	13	29	16	21	17	29	15	40	10	19	58
13	1	2	49	1	2	2	28	8	9	13	8	12	20	16	9	18	10	13	39
14	1	1	64	2	2	1	26	11	15	21	10	15	26	28	15	35	16	14	43
15	1	2	58	2	2	1	23	15	9	16	9	11	28	21	10	22	10	17	56
16	1	1	49	3	2	1	25	12	17	24	12	18	28	45	14	45	13	21	37
17	1	1	57	2	2	1	25	15	13	16	10	18	25	24	16	28	8	18	43
18	1	1	45	4	2	1	16	12	8	11	6	3	25	16	11	17	9	9	29
19	1	1	45	3	2	1	25	16	12	23	10	13	32	46	18	35	14	19	45
20	1	1	60	3	2	2	19	19	8	23	14	10	28	50	12	38	12	13	51
21	1	2	65	4	2	1	29	24	14	41	10	23	29	48	20	51	15	24	55
22	1	1	44	4	1	1	25	15	17	22	11	16	32	42	19	45	15	19	49
23	1	2	38	3	1	1	20	9	2	7	8	9	22	23	17	19	15	14	31
24	1	1	35	4	1	1	11	6	8	16	8	9	22	27	16	20	14	15	40
25	1	2	42	2	1	1	15	7	8	11	12	7	14	13	7	21	12	10	48
26	1	2	50	2	1	1	26	14	10	36	13	17	25	18	14	42	13	18	35
27	1	1	61	4	2	2	28	42	16	40	16	11	24	27	15	20	14	14	62
28	1	2	34	4	1	1	17	13	7	10	5	11	21	17	13	13	10	15	42
29	1	1	60	3	1	2	23	13	9	23	11	14	28	20	18	45	14	21	58
30	1	2	39	2	1	1	11	5	5	5	5	1	27	15	13	9	8	12	29
31	1	1	39	3	1	1	24	4	11	25	11	17	21	11	12	13	9	10	49
32	1	2	41	2	1	1	24	8	3	14	8	8	21	17	10	18	9	14	30
33	1	2	55	3	1	1	31	15	17	45	16	24	32	43	18	46	13	21	62
34	1	2	42	3	1	1	23	11	7	15	6	7	29	27	14	33	9	19	58
35	1	1	50	4	1	1	18	17	12	28	12	17	29	41	15	48	12	22	55
36	1	1	58	2	1	1	13	12	6	10	6	11	29	23	11	27	8	10	33
37	1	2	59	3	1	2	27	7	13	23	12	10	32	39	18	49	16	19	55
38	1	2	36	1	1	2	11	12	9	5	5	3	12	12	6	13	9	6	27
39	1	1	51	2	1	1	32	16	16	34	15	17	21	17	6	21	8	12	62

(Continued)

SESAME STREET DATA (Continued)

I D	S			A G E	V I E W T A T G	S E T I N G	V I E W N C	P R E B O D Y	P R E L E T	P R E F O R M	P R E N U M B	P R E R E L A T	P R E C L A S S I F	P O S T B O L E Y	P O S T T E R M	P O S T N U M B	P O S T R E L	P O S T C L A S S	P O S T E R I O R	P O S T E R I O R	
40	1	1	51	2	1	1	31	18	13	33	15	14	21	16	11	7	9	11	58		
41	1	1	48	3	1	1	13	14	8	8	10	11	21	22	10	28	8	19	34		
42	1	1	43	2	1	1	17	13	14	13	11	14	24	19	12	22	11	20	32		
43	1	2	35	3	1	2	23	12	8	9	5	5	29	11	8	9	11	11	32		
44	1	2	36	1	1	2	11	2	6	5	2	4	21	6	11	6	6	7	28		
45	1	2	39	2	1	2	20	18	6	4	4	6	19	8	11	22	10	10	29		
46	1	1	45	4	1	2	14	13	9	16	9	12	29	48	17	48	14	19	35		
47	1	1	58	3	1	2	30	38	15	45	14	18	32	48	19	46	14	23	67		
48	1	2	38	3	1	1	13	10	7	8	5	7	26	36	14	20	10	16	29		
49	1	2	57	4	1	1	26	15	11	22	10	15	24	20	18	28	12	18	35		
50	1	1	49	3	1	2	26	35	10	47	13	17	26	13	7	12	11	14	67		
51	1	1	55	1	2	2	24	11	10	18	8	10	20	10	14	23	10	10	39		
52	1	2	44	4	2	2	25	39	18	41	9	21	30	47	20	50	11	23	90		
53	1	1	56	1	2	2	13	11	6	15	10	11	15	13	9	13	10	6	46		
54	1	2	48	2	2	2	17	11	9	14	8	9	26	32	15	27	11	11	39		
55	1	1	50	2	2	2	16	10	8	9	7	6	21	15	17	17	12	15	34		
56	1	1	52	1	2	2	16	15	6	13	11	13	19	14	14	18	7	16	38		
57	1	2	51	2	2	1	24	14	10	20	10	17	28	21	17	36	19	17	34		
58	1	2	58	1	2	2	25	17	17	23	14	15	25	16	19	28	4	20	37		
59	1	2	48	3	2	2	13	10	10	13	9	7	27	15	14	23	11	12	33		
60	1	1	54	4	2	2	16	13	10	10	9	7	19	20	14	19	12	11	36		
61	2	1	52	4	2	2	20	35	15	21	8	20	27	48	19	47	15	22	49		
62	2	2	48	1	2	2	20	12	11	13	7	11	28	19	17	17	11	17	53		
63	2	1	55	4	2	2	28	13	10	29	11	12	30	31	19	45	15	22	65		
64	2	1	55	2	2	2	23	16	5	32	10	13	28	28	15	46	13	17	55		
65	2	2	55	2	1	2	30	27	11	39	12	17	32	40	19	52	17	23	85		
66	2	2	56	1	1	1	26	18	10	30	12	9	29	46	13	44	13	17	43		
67	2	2	50	3	2	1	31	15	10	24	11	20	31	43	18	52	14	23	65		
68	2	1	51	3	2	1	28	19	14	37	13	15	32	47	19	48	15	22	75		
69	2	1	58	3	2	2	30	14	17	37	13	20	28	38	15	36	16	18	85		
70	2	2	55	4	2	2	27	10	10	12	14	16	31	42	16	31	13	20	40		
71	2	2	41	4	1	2	24	22	14	42	14	21	30	49	19	50	14	22	58		
72	2	2	51	4	1	2	20	13	11	22	11	15	27	17	11	20	14	14	42		
73	2	1	52	4	2	1	29	13	9	18	10	10	30	23	17	28	12	17	69		
74	2	2	54	3	2	1	30	26	13	23	10	17	32	42	12	37	12	19	58		
75	2	1	47	4	1	1	19	12	11	16	11	12	28	43	18	38	14	17	37		
76	2	1	50	4	1	1	31	30	15	47	15	19	23	48	19	49	13	20	62		
77	2	2	55	4	1	1	31	13	18	39	15	24	31	51	19	50	14	23	99		
78	2	1	50	3	2	1	32	27	13	29	13	12	31	45	19	53	16	23	75		

(Continued)

SESAME STREET DATA

					V	S	V	P		P	P	P	P	P	P	P	P	P	P
					I	E	I	R	P	R	R	R	R	R	R	R	R	R	R
					E	T	E	E	E	E	E	E	E	E	E	E	E	E	E
					W	T	W	B	O	L	O	U	L	A	O	L	O	U	R
I	S	S	A	C	I	E	N	D	E	R	M	A	S	D	E	R	M	E	A
D	T	X	E	T	G	C	Y	T	M	B	T	F	Y	T	M	B	L	S	Y
79	2	2	57	2	2	1	31	19	14	36	13	11	32	50	17	47	15	22	82
80	2	2	55	4	1	2	20	12	9	16	11	18	30	30	14	18	12	16	47
81	2	2	55	4	1	2	26	14	8	24	13	12	30	45	19	43	15	24	62
82	2	1	50	3	2	1	30	44	15	45	12	11	32	53	19	52	15	23	67
83	2	1	52	3	2	1	26	13	11	34	12	12	30	45	17	43	13	21	40
84	2	2	45	2	2	1	28	12	9	16	7	8	28	21	14	32	9	17	41
85	2	2	52	3	2	1	24	14	8	18	8	7	28	43	12	41	13	15	56
86	2	1	53	4	1	1	26	17	15	32	13	16	27	37	15	31	14	23	73
87	2	2	53	4	2	2	29	10	17	23	13	15	32	51	19	48	16	21	85
88	2	2	53	2	2	2	23	12	12	15	11	17	28	20	15	19	8	19	59
89	2	2	56	3	1	2	28	29	11	43	13	22	31	32	15	40	15	21	58
90	2	2	54	4	2	2	32	46	15	48	13	21	32	51	19	50	16	23	92
91	2	1	50	2	2	1	22	17	8	18	12	14	25	16	14	32	14	17	69
92	2	1	50	3	1	1	29	25	14	35	15	20	26	36	16	31	8	15	56
93	2	2	53	4	1	1	25	17	12	30	13	17	29	40	17	40	13	22	78
94	2	1	45	4	1	1	21	16	12	15	11	17	29	36	19	29	10	16	58
95	2	2	56	4	1	1	32	22	15	32	11	13	31	46	20	51	14	24	78
96	2	1	53	3	1	1	31	14	16	29	11	17	32	43	19	42	13	22	67
97	2	2	46	4	1	1	22	28	14	20	5	15	32	42	13	29	15	18	69
98	2	2	46	2	1	2	30	18	14	23	11	11	29	33	13	36	8	16	53
99	2	1	50	3	1	2	18	13	8	14	11	12	19	23	11	31	12	17	55
100	2	2	47	1	1	2	17	10	5	11	8	7	18	19	9	8	10	13	53
101	2	1	56	2	2	1	27	11	15	22	13	14	30	47	20	45	15	24	67
102	2	2	46	3	1	1	27	13	13	22	12	13	28	36	17	46	11	20	62
103	2	2	45	4	1	1	21	23	14	27	13	20	31	48	19	44	11	24	59
104	2	2	46	4	1	1	19	17	4	19	10	13	28	36	16	29	10	15	46
105	2	2	47	3	1	1	31	9	14	24	11	16	32	42	17	42	11	20	58
106	2	1	52	2	2	1	26	16	12	30	14	16	27	29	20	41	15	21	92
107	2	2	52	4	2	1	24	15	12	30	14	14	31	45	18	49	13	23	48
108	2	1	56	2	1	1	21	22	12	25	12	15	29	37	14	46	14	18	65
109	2	1	48	4	1	1	28	22	13	19	11	8	32	48	18	43	15	19	67
110	2	1	49	3	1	1	25	16	13	18	15	15	27	48	13	45	15	18	59
111	2	2	55	4	2	1	32	8	13	23	11	17	29	35	18	36	11	19	39
112	2	2	45	3	1	1	22	15	12	20	9	14	25	21	15	21	9	16	47
113	2	1	45	3	1	1	32	16	14	30	11	17	25	26	17	39	13	19	51
114	2	2	48	4	1	1	31	6	8	13	7	13	29	32	15	23	7	16	43
115	2	1	58	1	1	2	19	14	12	23	11	10	28	15	10	34	11	7	65
116	3	1	55	2	1	1	20	16	7	14	9	9	21	11	13	20	13	13	39
117	3	2	48	1	1	2	20	15	5	13	8	7	21	14	3	17	10	7	33

(Continued)

SESAME STREET DATA (Continued)

I D	S				V I E W C A T N G	S E T W E N C	V I E W E N C	P R E B O D Y	P R E L E T	P R E F O R M	P R E N U M B	P R E R E L A T	P R E C L A S S I F Y	P O S T E R I O R I T Y	P O S T E R I O R I T Y	P O S T E R I O R I T Y	P O S T E R I O R I T Y	P O S T E R I O R I T Y	P O S T E R I O R I T Y	P O S T E R I O R I T Y
118	3	2	52	3	1	1	14	6	3	9	4	8	18	19	20	18	12	20	34	
119	3	2	58	1	1	1	20	11	5	25	9	7	26	16	10	23	12	12	35	
120	3	1	50	3	2	2	13	12	5	11	10	9	21	18	10	19	10	9	32	
121	3	1	58	3	2	2	22	19	11	35	10	17	23	44	18	46	13	19	44	
122	3	2	49	4	2	2	14	13	7	5	5	7	17	16	12	19	12	15	31	
123	3	1	56	4	2	2	24	17	9	16	11	10	29	35	17	40	14	19	44	
124	3	1	50	1	1	2	7	14	4	10	5	5	19	15	13	14	13	17	27	
125	3	2	49	1	1	1	20	18	3	14	6	6	11	12	6	8	9	5	29	
126	3	1	46	2	1	1	15	14	8	23	14	12	25	18	14	30	12	16	47	
127	3	1	57	2	1	1	26	14	9	23	15	9	28	15	14	24	12	9	35	
128	3	1	44	1	2	2	12	9	7	14	9	14	13	13	7	17	9	7	32	
129	3	2	41	2	2	2	16	14	9	10	11	9	22	17	9	23	8	13	36	
130	3	2	58	4	2	1	17	9	11	28	8	13	29	13	12	29	12	15	38	
131	3	2	60	4	2	1	31	19	11	27	11	16	31	31	17	38	13	22	42	
132	3	2	40	2	1	1	12	14	3	17	6	6	16	13	6	17	10	9	27	
133	3	2	37	2	1	1	7	4	6	4	5	4	13	13	6	14	9	11	60	
134	3	1	45	1	1	1	12	5	5	9	7	7	15	13	12	20	12	11	28	
135	3	1	60	3	1	1	17	18	9	14	7	6	32	36	13	32	13	12	33	
136	3	1	52	2	1	2	18	13	9	24	10	16	25	15	12	26	11	10	55	
137	3	2	46	4	1	1	20	12	4	17	8	8	28	22	17	38	14	20	29	
138	3	2	60	4	1	1	23	16	9	25	11	14	29	26	17	38	16	22	46	
139	3	1	60	3	1	1	17	11	10	15	10	14	11	13	7	16	9	13	33	
140	3	1	59	3	1	1	7	16	11	10	6	10	15	14	9	14	8	10	32	
141	3	2	52	3	1	1	29	20	8	37	13	13	28	46	12	42	13	15	47	
142	3	1	60	3	1	1	29	13	12	17	12	16	29	25	17	32	13	19	90	
143	3	2	56	2	1	1	21	12	12	17	9	14	23	26	16	34	10	20	61	
144	3	2	54	2	2	2	18	28	9	14	12	16	27	42	18	37	11	15	36	
145	3	2	61	3	1	1	13	12	8	16	7	11	28	15	15	18	7	15	35	
146	3	2	61	3	1	1	29	18	12	22	11	14	30	25	17	39	13	19	48	
147	3	2	51	3	1	1	17	15	5	11	10	11	32	43	14	44	15	21	35	
148	3	1	49	4	2	1	19	17	7	16	3	6	27	27	18	43	15	20	35	
149	3	1	52	2	1	1	22	13	10	20	11	14	22	14	9	21	9	7	35	
150	3	2	55	3	2	1	25	13	12	16	10	14	26	17	6	31	13	9	35	
151	3	2	60	4	2	1	28	10	10	22	12	15	28	15	15	30	11	17	45	
152	3	2	43	1	2	2	14	9	5	15	5	6	16	16	12	14	10	14	33	
153	3	1	55	3	2	1	14	7	9	15	9	12	18	15	16	22	11	16	42	
154	3	2	52	4	2	1	18	11	9	15	8	12	23	23	15	40	13	18	32	
155	3	2	56	1	2	1	26	24	13	25	9	10	17	7	3	13	6	5	40	
156	3	1	56	4	1	1	24	11	11	28	14	17	27	14	15	40	12	21	42	

(Continued)

SESAME STREET DATA (Continued)

				V I E W I E O D Y T M B T F Y T M B L S Y				S E T W E B L O U L A S D E R M E A S Y				P R E R C T S T F N T C L O				P O S S O S T C L S			
I D	S I T E	S E X	A G E	V I E W	S E T	V I E W	P R E B O D Y	P R E F O R M	P R E N U M B	P R E L A T	P R E C L A S S	P O S T B O D Y	P O S T F O R M	P O S T N U M B	P O S T U R E	P O S T C L A S S	P O S T E A B O D Y		
157	3	2	47	2	1	1	20	19	9	25	12	8	26	24	13	35	11	17	46
158	3	2	56	2	1	1	17	18	8	17	5	9	24	17	10	19	10	15	39
159	3	2	52	3	1	1	28	15	13	27	9	15	31	16	16	22	12	18	69
160	3	2	51	4	1	1	23	14	11	23	8	11	31	37	17	42	14	18	36
161	3	1	51	1	1	1	7	13	6	11	7	6	12	8	14	22	10	16	35
162	3	2	53	3	1	1	15	15	8	18	8	11	29	32	12	28	10	14	34
163	3	1	50	4	1	1	26	11	14	23	10	11	39	22	16	40	14	17	32
164	3	2	59	4	1	1	16	10	8	21	9	12	25	22	14	31	10	18	38
165	3	1	53	3	1	1	14	12	7	9	9	5	22	28	9	30	9	10	32
166	3	1	55	3	1	1	15	10	7	9	6	11	24	20	14	27	9	9	34
167	3	1	57	1	1	1	6	13	2	8	7	7	18	6	4	0	1	4	35
168	3	1	58	2	1	1	16	5	5	8	6	9	13	14	11	11	9	16	34
169	3	1	44	3	1	1	10	12	4	9	10	11	13	15	3	8	3	5	28
170	3	1	39	1	1	2	14	12	4	5	7	5	13	11	8	19	10	8	29
171	3	1	53	4	2	1	21	17	12	16	10	13	27	20	14	29	15	16	37
172	3	2	52	4	1	1	23	10	9	9	7	6	21	16	11	20	9	9	32
173	3	1	57	3	1	2	25	11	10	19	11	13	28	29	20	25	16	19	35
174	3	2	40	3	1	1	11	10	7	14	4	8	16	22	11	21	9	9	35
175	3	2	47	2	1	1	16	13	7	7	6	9	22	13	4	18	11	9	32
176	3	1	51	2	1	1	25	19	11	24	12	8	26	20	15	24	11	13	47
177	3	1	48	2	1	1	11	7	4	14	3	13	11	12	8	27	11	10	35
178	3	2	49	1	1	2	15	16	6	9	4	7	20	16	7	17	10	5	35
179	3	1	50	2	1	1	12	8	5	17	8	10	18	19	12	13	12	17	30
180	4	2	53	1	2	2	10	13	4	13	7	8	19	16	9	16	7	11	35
181	4	2	52	1	2	2	13	15	8	19	8	9	21	11	8	16	7	11	39
182	4	1	51	1	2	2	19	12	9	17	8	12	27	16	12	27	11	16	39
183	4	1	52	1	2	2	20	16	12	22	11	17	25	19	14	26	11	15	36
184	4	1	46	1	2	2	13	3	3	1	4	4	24	11	10	13	11	13	27
185	4	2	51	1	2	2	21	19	12	25	13	14	24	15	11	25	8	14	45
186	4	2	47	1	2	2	19	12	13	27	8	11	24	14	15	21	7	13	28
187	4	2	51	3	2	1	25	13	12	21	12	16	31	16	15	25	11	18	40
188	4	2	54	1	2	1	8	20	5	8	7	6	14	13	11	11	8	10	47
189	4	2	54	2	2	1	12	4	9	4	7	6	17	13	10	12	9	8	36
190	4	1	57	2	2	1	24	11	10	28	12	11	30	24	18	26	14	16	39
191	4	1	53	2	2	1	17	12	8	9	5	11	26	13	11	20	10	15	39
192	4	2	50	2	2	1	20	16	8	18	9	13	28	25	15	15	9	12	43
193	4	2	57	1	2	2	28	23	16	33	14	11	26	25	16	42	14	11	69
194	4	2	58	1	2	2	31	30	12	44	14	17	32	43	16	44	11	13	69
195	4	2	58	1	2	2	28	29	9	33	14	8	29	44	9	44	15	10	38

(Continued)

SESAME STREET DATA (Continued)

					V	S	V	P		P	P	P	P	P	P	P	P	P
					I	E	I	R	P	R	R	R	R	R	R	R	R	R
					E	T	E	E	R	E	E	E	E	E	E	E	E	E
	S				W	T	W	B	E	F	N	E	L	B	T	F	N	T
	I	S	A		C	I	E	O	L	O	U	L	A	O	L	O	U	R
	T	E	G		A	N	N	D	E	R	M	A	S	D	E	R	M	E
	D	X	E		T	G	C	Y	T	M	B	T	F	Y	T	M	B	L
196	4	2	53	1	2	2	19	19	14	24	11	16	21	13	9	31	10	16
197	4	2	49	1	2	2	20	17	7	13	9	10	30	15	6	21	10	9
198	4	2	51	1	2	2	10	1	2	2	4	0	13	0	0	0	0	34
199	4	1	58	1	2	2	22	13	9	13	10	9	18	18	11	13	11	8
200	4	2	51	1	2	2	18	12	4	10	5	9	17	10	8	14	5	10
201	4	2	53	1	2	2	21	17	9	18	9	11	28	15	9	19	12	9
202	4	2	56	3	2	1	29	17	17	32	10	20	30	33	17	38	12	20
203	4	2	51	3	1	1	19	11	10	19	8	7	22	19	11	39	11	21
204	4	1	47	1	1	1	23	12	11	14	11	13	29	15	13	22	14	16
205	4	2	54	4	1	1	23	14	12	23	8	15	28	41	16	35	16	22
206	4	1	54	4	1	1	17	15	6	15	4	11	24	30	13	42	17	20
207	4	2	46	1	1	1	22	14	7	15	3	14	29	24	18	36	13	23
208	4	2	52	2	1	1	20	15	14	19	5	13	27	45	16	38	17	22
209	4	2	48	1	1	1	24	18	5	21	9	11	23	17	10	16	15	9
210	4	1	49	2	1	2	17	21	7	23	9	4	13	14	13	35	15	13
211	4	1	58	1	1	2	14	7	3	17	13	6	22	15	11	23	13	9
212	4	1	46	3	1	2	18	13	10	11	7	9	22	14	13	23	10	13
213	4	1	57	1	1	2	27	19	11	20	10	15	27	19	8	29	23	11
214	4	1	48	4	1	2	27	12	15	23	11	16	27	17	13	27	13	10
215	4	2	52	2	1	2	23	8	9	16	12	8	20	16	13	23	10	12
216	4	1	57	2	1	1	29	17	12	24	12	16	31	32	12	17	9	10
217	4	1	46	3	1	1	18	9	10	12	9	14	24	18	11	13	8	10
218	4	1	55	2	1	1	14	12	8	14	11	11	27	40	18	35	16	21
219	4	1	44	3	1	1	8	11	6	10	6	9	31	23	15	18	10	13
220	4	1	56	1	2	2	26	15	12	29	11	12	25	23	13	40	12	16
221	4	2	44	2	1	1	14	14	12	10	10	12	25	23	16	26	13	12
222	4	1	59	4	1	1	28	17	12	27	10	11	31	46	19	29	16	20
223	5	1	48	2	1	1	16	8	8	9	8	8	24	11	9	11	8	7
224	5	1	56	2	1	1	22	17	11	23	14	13	30	20	17	38	13	21
225	5	2	58	2	1	1	20	18	8	26	11	10	30	44	12	40	13	21
226	5	2	53	1	1	1	15	11	2	8	10	5	18	19	10	14	6	8
227	5	2	53	1	1	1	26	16	8	14	10	9	28	13	12	18	10	11
228	5	2	65	1	1	2	15	16	5	24	12	12	22	15	12	26	11	19
229	5	1	46	1	1	2	15	5	5	4	9	4	15	13	10	10	8	11
230	5	1	49	1	1	2	19	12	12	16	13	15	28	16	14	36	14	16
231	5	1	55	1	1	2	21	40	8	36	9	10	27	49	13	47	9	17
232	5	2	46	2	1	1	20	9	6	17	10	11	29	13	12	23	12	8
233	5	2	58	4	1	1	30	55	19	52	15	23	31	54	19	54	15	23

(Continued)

SESAME STREET DATA (Continued)

				V S V P				P P P P				P P P P				P P P P			
				I E I R	P R R E	R E R C	T S T S	O S O S	O S O S	O S O S	O S O S	O S O S	O S O S	O S O S	O S O S	O S O S	O S O S	O S O S	O S O S
				W T W B	E R E F	N E N L	B T B T	F N F N	L B L B	O L O L	O L O L	O L O L	O L O L	O L O L	O L O L	O L O L	O L O L	O L O L	O L O L
				A C I E	O O D E	R M A S	D E R M	E R E M	E R E M	E R E M	E R E M	E R E M	E R E M	E R E M	E R E M	E R E M	E R E M	E R E M	E R E M
I	T	E	G	A	N	N	D	E	R	M	A	S	D	E	R	M	E	A	D
D	E	X	E	T	G	C	Y	T	M	B	T	F	Y	T	M	B	L	S	Y
234	5	1	47	4	1	1	18	13	9	20	8	11	28	34	17	33	11	18	43
235	5	1	53	4	1	1	26	25	14	36	13	13	30	44	19	43	15	23	90
236	5	2	51	2	1	1	30	15	8	12	10	10	30	33	12	45	12	20	49
237	5	1	49	4	1	1	17	16	12	15	8	15	25	26	15	20	12	11	41
238	5	1	43	2	1	1	16	13	6	11	8	9	22	19	10	10	9	7	30
239	5	2	60	3	1	1	23	16	9	33	14	16	29	35	18	50	13	23	69
240	5	1	51	4	1	1	21	11	10	27	10	12	25	32	17	47	11	19	65

DESCRIPTION OF HEADACHE DATA SET

This study investigated the effectiveness of different kinds of psychological treatment on the sensitivity of headache sufferers to noise. Each subject was exposed to the following sequence of operations: (1) measurement of initial sensitivity scores, (2) relaxation training, (3) treatment, and (4) measurement of final sensitivity scores.

The sensitivity scores were obtained by having subjects listen to a tone that gradually increased in volume and asking them to rate the levels at which the tone became (1) uncomfortable and (2) definitely unpleasant. These levels, denoted by U and DU, are the dependent variables, with pretest scores on these variables useful for possible covariance analysis.

Relaxation training was applied to all subjects and comprised two stages: (1) The subjects were asked to listen to the tone at their definitely unpleasant level for up to two minutes (with the option to terminate the exposure if they chose). (2) The subjects were then given instruction on breathing techniques and the use of visual imagery to act as a controlled distraction.

There were two types of headache sufferers in the study: (a) migraine and (b) tension. Within each of these groups subjects were randomly assigned to one of the following four treatment groups:

- T1— subjects in this group listened to the tone again at their definitely unpleasant (DU) level for the length of time that they were able to stand it in (a) above.
- T2— as T1 but with one extra minute's exposure to the tone.
- T3— as T2 but having been instructed to use the relaxation techniques of breathing and imagery.
- T4— this was a control group, in that the subjects experienced no exposure to the tone between (a) in the relaxation training and the final sensitivity measures.

Some missing data reduced an intended balanced design to the following 2 × 4 factorial design, with cell sizes indicated:

	T1	T2	T3	T4
MIGRANE	11	11	12	11
TENSION	14	11	16	12

HEADACHE DATA

H E A D A C H E T Y P E	T R A T T M E N T	P R E U N C I P A L	P R E U N C I P A L	U N C E R T A I N	D I F F I C U L T	H E T Y P E	T R A T T M E N T	P R E U N C I P A L	P R E U N C I P A L	U N C E R T A I N	D I F F I C U L T
1	3	2.34	5.30	5.80	8.52	1	1	2.73	6.85	4.68	6.68
2	1	0.37	0.53	0.55	0.84	1	3	7.50	9.12	5.70	7.88
1	3	4.63	7.21	5.63	6.75	1	3	3.60	7.30	4.83	7.32
1	2	2.45	3.75	2.50	3.18	1	1	2.31	3.25	2.00	3.30
1	1	1.38	2.33	2.23	3.98	2	3	0.85	1.42	1.37	1.89
1	3	1.85	3.25	3.40	4.80	1	2	1.90	8.68	2.25	6.70
2	1	6.00	9.90	8.25	10.7	1	2	1.56	2.92	2.00	2.84
2	1	2.95	4.98	3.85	4.75	2	1	2.95	3.45	1.75	2.30
2	2	6.68	9.90	8.52	12.8	1	4	1.72	2.75	2.20	3.95
2	3	3.90	6.50	3.27	7.80	1	3	0.40	0.90	1.40	2.30
1	1	2.19	2.60	2.50	3.50	2	4	1.40	1.82	2.10	3.90
2	3	3.22	5.65	2.70	4.80	2	3	3.50	6.60	4.65	8.00
2	2	3.15	5.25	5.30	7.60	1	4	1.96	3.18	1.20	3.15
2	2	2.55	4.05	4.00	5.45	2	1	1.85	3.30	1.80	3.15
2	4	1.85	3.20	1.42	2.62	2	1	1.50	1.75	1.35	3.40
2	4	4.32	6.15	4.98	6.45	2	1	2.43	7.95	4.08	6.83
1	2	3.42	5.59	4.50	7.18	2	3	3.70	5.88	3.13	4.00
2	3	2.57	4.40	3.27	8.64	1	3	1.62	3.40	4.03	5.70
2	2	4.66	6.82	3.45	6.24	1	2	1.12	1.39	1.06	1.78
2	4	4.10	7.65	3.36	6.58	1	4	2.65	4.88	1.20	3.50
1	4	0.66	1.00	0.43	0.60	2	4	2.08	3.30	2.44	3.47
2	2	3.38	8.27	7.07	0.90	2	4	3.86	5.94	3.20	4.81
2	2	2.39	4.60	2.93	5.42	2	3	3.62	8.83	5.12	7.71
1	4	1.86	4.06	1.78	2.44	2	3	2.19	3.94	2.31	3.48
2	4	2.08	2.64	1.71	2.99	2	4	1.51	2.80	1.24	2.63
1	2	1.60	2.83	1.87	3.00	1	3	0.75	0.94	0.88	1.45
1	1	0.87	1.16	0.59	0.95	1	3	1.92	2.44	2.00	2.54
2	4	1.42	6.47	2.00	3.48	1	4	1.82	2.57	0.64	1.07
2	1	1.86	2.74	0.89	1.41	1	4	1.24	2.69	0.95	1.76
1	3	0.90	1.41	1.56	2.11	2	3	1.24	3.23	1.36	2.86
1	4	1.51	2.79	0.83	1.64	1	1	1.56	8.69	2.35	7.51
2	3	1.44	3.06	1.11	2.58	2	2	0.46	1.12	0.93	1.36
1	2	2.20	5.25	1.04	3.19	2	2	1.29	2.32	1.56	5.30

(Continued)

HEADACHE DATA (Continued)

H E A D A C H E T Y P E		P R E U N C O M F	P R E D F U N C O M P L			H E A D A C H E T Y P E		P R E U N C O M F	P R E D F U N C O M P L		
2	2	2.34	4.25	2.16	4.10	1	1	0.86	1.55	0.88	2.14
1	2	5.91	8.56	2.62	6.08	2	2	1.43	3.94	3.61	7.57
2	3	15.1	16.3	15.5	16.4	2	1	0.55	1.10	1.80	3.92
2	3	7.42	14.5	8.15	13.3	1	4	3.40	5.10	2.80	4.40
2	3	1.52	2.35	1.20	2.55	2	1	1.85	5.68	7.75	16.1
1	1	2.25	4.40	1.56	4.93	1	2	4.22	13.3	12.2	14.1
2	3	3.30	4.55	5.25	5.83	2	3	4.30	11.3	8.78	12.38
1	3	3.58	5.60	6.94	9.16	2	3	6.17	15.5	7.54	16.24
2	1	0.43	0.64	0.22	0.39	1	1	1.33	10.3	1.67	3.79
2	1	0.87	1.30	1.10	1.45	1	2	5.05	10.05	3.02	10.1
2	1	1.88	4.19	1.79	4.26	1	1	8.94	15.46	3.64	9.00
1	1	2.50	4.64	2.23	3.60	2	1	13.3	17.00	11.87	16.6
2	4	3.25	5.09	1.24	2.11	1	4	10.36	17.0	11.50	15.56
1	2	1.08	1.49	1.09	1.82	2	4	3.01	12.35	4.01	7.51
1	4	0.71	1.13	0.58	0.86	2	2	2.00	8.37	10.08	5.49
1	3	12.72	16.5	15.2	16.8	2	4	11.3	16.8	6.75	16.87

DESCRIPTION OF THE CARTOON DATA SET

This is a data set on 179 subjects from the *Minitab Handbook* (2nd ed., 1985), and is used with permission of the publisher. A short instructional slide presentation was developed, which dealt with the behavior of people in a group situation, and in particular the various roles or character types that group members often assume. The presentation consisted of a 5 minute lecture on tape, accompanied by 18 slides. Each role was identified by an animal. Each animal was shown on two slides: once in a cartoon sketch and once in a realistic picture. All 179 subjects saw all 18 slides, but a randomly selected half of them saw the slides in black and white while the other half saw the slides in color.

After seeing the slides, the subjects took a test on the material. The slides were presented in random order, and the subjects wrote down the character type represented by that slide. They received two scores: one for the number of cartoon characters correctly identified and one for the number of realistic characters correctly identified. Each score could range from 0 to 9, since there were 9 characters. Four weeks later the subjects were retested. Some subjects did not show up for the retest and that is indicated by a blank.

There are three groups of subjects in this study: (1) preprofessional personnel at three hospitals in Pennsylvania involved in an in-service training program, (2) professional personnel involved in the same training program, and (3) a group of Penn State undergraduate students. All these subjects were given the Otis Mental Ability Test, which yields a rough estimate of their natural ability.

The order in which the variables are arranged on the file is as follows:

<i>Variable No.</i>	<i>Variable Name</i>	<i>Description</i>
1	ID	Identification number
2	COLOR	0 = black and white, 1 = color (no participant saw both)
3	ED	Education: 0 = preprofessional, 1 = professional, 2 = college student
4	LOCATION	Location: 1 = hospital A, 2 = hospital B, 3 = hospital C, 4 = Penn State student
5	OTIS	OTIS score: from about 70 to about 130
6	CARTOON 1	Score on cartoon test given immediately after presentation (possible scores are 0,1,2,. . .9)
7	REAL1	Score on realistic test given immediately after presentation (possible scores are 0,1,2,. . .9)
8	CARTOON2	Score on cartoon test given four weeks (delayed) after presentation (possible scores are 0,1,2,. . .9; a blank is used for a missing observation)
9	REAL2	Score on realistic test given four weeks (delayed) after presentation (possible scores are 0,1,2,. . .9; a blank is used for a missing observation)

CARTOON DATA

				C				C								C				C			
				A				A								A				A			
				R				R								R				R			
				T				T								T				T			
				O				O								O				O			
				E				E								E				E			
				L				L								L				L			
				D				D								D				D			
				U				U								U				U			
				O				O								O				O			
				I				I								I				I			
				N				N								N				N			
				S				S								S				S			
				1				1								1				1			
				2				2								2				2			
				2				2								2				2			
1	0	0	1	107	4	4						40	1	0	3	108	9	9					
2	0	0	2	106	9	9	6	5				41	1	0	3	86	4	4					
3	0	0	2	94	4	2	3	0				42	1	0	3	96	6	3					
4	0	0	2	121	8	8	6	8				43	1	0	3	101	5	3					
5	0	0	3	86	5	5						44	1	0	3	97	6	3	4	4			
6	0	0	3	99	7	8	7	5				45	1	0	3	88	3	1	2	0			
7	0	0	3	114	8	9	5	4				46	1	0	3	104	4	2	2	0			
8	0	0	3	100	2	1						47	1	0	3	87	7	3					
9	0	0	3	85	3	2						48	1	0	3	86	1	1					
10	0	0	3	115	8	7	8	5				49	1	0	3	90	6	5	4	1			
11	0	0	3	101	7	6						50	1	0	3	102	6	2					
12	0	0	3	84	7	5						51	1	0	3	105	2	2					
13	0	0	3	94	4	3						52	1	0	3	115	7	8					
14	0	0	3	87	1	3	2	0				53	1	0	3	88	4	3					
15	0	0	3	104	9	9	5	6				54	1	0	3	111	8	8					
16	0	0	3	104	5	6						55	1	0	3	95	5	4					
17	0	0	3	97	6	5						56	1	0	3	104	5	5					
18	0	0	3	91	1	0						57	0	1	1	79	7	4	6	4			
19	0	0	3	83	4	4						58	0	1	1	82	3	2					
20	0	0	3	93	0	0						59	0	1	1	123	8	8	7	5			
21	0	0	3	92	2	2						60	0	1	1	106	9	7	8	6			
22	0	0	3	91	5	2	3	1				61	0	1	1	125	9	9	4	3			
23	0	0	3	88	2	1						62	0	1	1	98	7	6					
24	0	0	3	90	5	4	4	3				63	0	1	1	95	7	7	4	4			
25	0	0	3	103	6	2						64	0	1	2	129	9	9	7	7			
26	0	0	3	93	9	9	8	4				65	0	1	2	90	7	6	3	5			
27	0	0	3	106	2	0	6	3				66	0	1	2	111	6	2	3	1			
28	1	0	1	98	3	3						67	0	1	2	99	4	5	3	1			
29	1	0	1	103	6	5	2	2				68	0	1	2	116	9	7	7	7			
30	1	0	2	109	5	4	1	2				69	0	1	2	106	8	7	6	4			
31	1	0	2	107	8	8						70	0	1	2	107	8	5					
32	1	0	2	108	8	8	7	6				71	0	1	2	100	7	6	2	1			
33	1	0	2	107	3	2						72	0	1	2	124	8	9	3	5			
34	1	0	3	87	6	4	2	2				73	0	1	3	98	6	7	1	1			
35	1	0	3	113	5	4	4	4				74	0	1	3	124	9	6	6	5			
36	1	0	3	80	0	3	1	1				75	0	1	3	84	1	4					
37	1	0	3	91	5	6						76	0	1	3	91	8	3					
38	1	0	3	102	8	9	5	5				77	0	1	3	118	6	6	3	4			
39	1	0	3	83	4	1	2	1				78	0	1	3	102	6	4					

(Continued)

CARTOON DATA (Continued)

				C A R T				C A R T								C A R T				C A R T			
				O	E	O	R	O	E	O	R	O	E	O	R	O	E	O	R	O	E	O	R
				L	D	L	T	O	A	O	A	L	D	L	T	O	A	O	A	L	D	L	T
I	O	U	O	I	N	L	N	L	I	O	U	O	I	N	L	N	L	I	O	U	O	I	N
D	R	C	C	S	1	1	2	2	D	R	C	C	S	1	1	2	2	D	R	C	C	S	1
79	0	1	3	95	7	4			118	0	2	4	97	8	8	6	4						
80	0	1	3	90	4	3			119	0	2	4	123	9	9	7	4						
81	0	1	3	86	1	0			120	0	2	4	113	8	7	6	6						
82	0	1	3	104	6	4			121	0	2	4	110	8	7	3	5						
83	1	1	1	111	9	9	6	3	122	0	2	4	119	8	7	6	6						
84	1	1	1	105	1	0			123	0	2	4	116	5	7								
85	1	1	1	110	1	0	0	0	124	0	2	4	113	8	6	5	5						
86	1	1	1	80	0	0	0	0	125	0	2	4	128	9	9								
87	1	1	1	78	4	1	1	1	126	0	2	4	113	8	5	4	2						
88	1	1	2	120	9	9			127	0	2	4	110	5	7								
89	1	1	2	110	9	6	6	5	128	0	2	4	114	7	6	5	5						
90	1	1	2	107	8	6			129	0	2	4	132	9	8	4	6						
91	1	1	2	125	7	8			130	0	2	4	110	7	8	2	5						
92	1	1	2	117	9	9			131	0	2	4	122	7	7	4	2						
93	1	1	2	126	8	8	5	5	132	0	2	4	123	9	9	6	7						
94	1	1	2	98	4	5			133	0	2	4	131	9	9	7	7						
95	1	1	2	111	8	6			134	0	2	4	131	9	9	8	8						
96	1	1	2	110	8	7			135	0	2	4	121	9	8	7	8						
97	1	1	2	120	9	7			136	0	2	4	125	9	8								
98	1	1	2	114	8	7	6	4	137	0	2	4	101	6	6	4	6						
99	1	1	2	117	6	7			138	0	2	4	120	8	9	6	7						
100	1	1	3	105	7	6			139	0	2	4	99	9	6								
101	1	1	3	97	6	6			140	0	2	4	128	8	9	8	7						
102	1	1	3	86	1	1			141	0	2	4	129	8	6	5	2						
103	1	1	3	111	7	5			142	0	2	4	125	8	6	7	4						
104	1	1	3	93	1	0			143	0	2	4	107	8	8	8	5						
105	1	1	3	115	8	7			144	0	2	4	102	8	7	6	4						
106	1	1	3	102	2	3	5	2	145	0	2	4	125	9	8								
107	1	1	3	111	7	3	4	4	146	1	2	4	129	8	8								
108	1	1	3	82	1	1			147	1	2	4	122	3	0	2	3						
109	1	1	3	117	8	5	4	3	148	1	2	4	124	7	6	6	7						
110	0	2	4	132	9	9			149	1	2	4	115	8	8								
111	0	2	4	113	7	8			150	1	2	4	117	8	6	5	2						
112	0	2	4	130	9	7	1	4	151	1	2	4	132	7	6	5	7						
113	0	2	4	122	9	9	6	4	152	1	2	4	109	8	5	5	5						
114	0	2	4	103	7	5	3	0	153	1	2	4	107	9	5	9	2						
115	0	2	4	103	7	5	3	0	154	1	2	4	116	8	7	6	5						
116	0	2	4	118	9	9			155	1	2	4	118	8	5	6	5						
117	0	2	4	119	9	9	7	8	156	1	2	4	124	9	9	6	7						

(Continued)

DESCRIPTION OF ATTITUDE DATA SET

Data was collected on 189 third through sixth graders from a suburban mid-western public school. The children were measured on preference toward the following subjects: mathematics, language arts, science, reading, and social studies. An intervention was then employed with the teachers (five) to change these preferences (attitudes) in a positive way, and then the children were measured again on subject preference four months later.

The intervention consisted of a three hour lecture and discussion (in-service work) by a professor in September with the teachers on what shapes attitudes and how they could go about changing the attitudes of their students. There was a particular emphasis in this school on changing the mathematics attitude. The professor met again with the teachers in December to discuss whether they had implemented some of the changes he had suggested.

ATTITUDE DATA

P P P P P													P P P P P												
T P P P P P O O O O O													T P P P P P O O O O O												
E R R R R R S S S S S													E R R R R R S S S S S												
G A E E E E T T T T T													G A E E E E T T T T T												
R C M L S R S M L S R S													R C M L S R S M L S R S												
S A H A A C E O A A C E O													S A J A A C E O A A C E O												
E D E T N I A C T N I A C													E D H T N I A C T N I A C												
X E R H G E D S H G E D S													X E R H G E D S H G E D S												
1	3	5	1	4	2	5	4	0	5	1	4	2	1	3	4	6	4	3	4	2	6	3	3	2	1
2	3	5	4	7	3	4	5	1	7	4	4	5	1	3	4	7	4	3	5	1	5	2	3	4	3
1	3	5	8	6	5	3	7	8	5	4	3	5	1	3	4	3	0	6	4	1	2	0	6	4	2
2	3	5	1	7	2	4	3	4	6	1	5	2	1	3	4	1	1	7	4	5	2	2	6	3	1
2	3	5	7	3	1	7	5	5	2	5	4	6	2	3	4	4	2	1	5	2	3	2	4	2	4
2	3	5	3	6	3	7	4	1	5	4	2	4	2	3	4	6	4	1	4	2	6	2	6	3	4
1	3	5	5	3	1	6	2	5	2	2	7	4	1	3	4	6	6	5	4	3	6	1	8	5	7
1	3	5	6	1	5	7	3	7	2	5	3	6	1	3	4	2	1	3	6	3	4	5	2	7	2
2	3	5	7	4	3	5	1	5	2	4	6	3	2	3	4	3	2	3	6	1	4	2	3	4	3
2	3	4	3	1	2	4	3	5	2	1	2	3	1	3	4	5	3	3	2	2	7	2	6	2	5
2	3	4	7	2	4	4	3	5	5	2	6	1	2	6	2	5	1	5	7	5	3	2	5	7	0
2	3	4	6	1	3	4	2	3	1	5	7	2	1	6	2	2	1	6	0	7	3	0	7	1	7
1	3	4	7	1	6	5	1	6	1	6	5	2	2	6	2	5	4	2	2	7	4	3	4	5	4
1	3	4	7	2	3	4	2	7	0	4	6	3	1	6	2	4	0	4	1	7	7	2	4	3	6
1	3	4	6	3	1	3	5	6	1	7	2	5	2	6	2	4	3	2	4	3	4	1	5	1	2
2	3	4	4	4	1	4	2	5	2	2	7	3	1	6	2	3	1	2	4	7	8	2	1	3	5
2	3	4	6	2	3	4	1	7	1	4	4	0	1	6	2	6	1	3	2	5	7	3	5	6	2
2	3	4	5	1	4	5	5	4	4	3	6	1	2	6	3	3	6	8	4	7	3	8	4	4	5
2	3	4	6	5	1	7	3	4	5	1	3	3	1	6	2	6	3	5	4	7	8	4	5	1	6
2	3	4	4	2	1	6	0	4	2	1	7	0	1	6	2	6	1	5	3	8	8	3	5	4	6
1	3	4	5	6	2	5	4	7	4	2	6	5	2	6	2	1	5	4	3	5	2	6	3	5	2
2	3	4	6	2	1	6	4	6	1	4	5	2	1	6	2	7	4	8	2	6	0	1	3	4	5
1	3	4	5	2	2	4	2	7	2	1	7	2	1	6	2	2	0	4	4	5	8	4	7	0	5
2	3	4	4	2	6	6	0	4	3	4	6	3	1	6	2	8	3	5	5	7	6	2	4	5	1
2	6	2	0	6	3	5	4	0	2	3	4	5	1	6	2	6	2	5	4	4	4	4	7	4	7
1	6	2	0	3	7	1	5	0	1	5	2	3	2	6	2	5	7	3	2	5	4	5	1	2	3
2	6	2	6	2	8	5	4	7	2	8	5	3	2	6	2	4	2	7	1	6	5	2	8	0	5
2	6	2	5	3	6	3	7	3	5	7	0	6	2	4	1	7	2	6	4	0	6	2	7	2	1
2	6	2	4	6	2	8	1	3	6	4	8	2	2	4	1	8	6	1	5	0	6	5	1	3	0
1	6	2	4	2	5	7	7	1	3	4	7	8	1	4	1	6	0	5	1	2	6	2	8	0	3
2	6	2	2	5	8	5	5	5	4	7	2	6	2	4	1	4	5	2	6	2	4	3	4	5	0
2	6	2	8	1	3	1	2	7	4	2	5	0	2	4	1	7	4	3	7	2	7	4	2	7	0
1	6	2	6	0	8	4	4	8	2	7	3	4	2	4	1	7	5	1	7	2	5	2	2	5	0
2	6	2	7	5	7	1	7	6	4	8	1	7	1	4	1	7	1	5	5	8	4	2	5	6	6
2	6	2	6	2	3	5	3	6	3	1	5	2	1	4	1	7	4	4	5	2	7	5	1	6	3
1	6	2	3	4	7	2	5	6	3	8	4	7	2	4	1	6	5	3	8	1	5	3	5	8	0
2	6	2	4	1	6	2	7	4	0	7	6	1	2	4	1	8	4	1	7	0	8	6	5	7	3
2	6	2	7	6	5	2	3	5	6	7	3	3	1	4	1	6	2	6	6	3	5	1	8	2	4

(Continued)

ATTITUDE DATA (Continued)

S E X	P O P P P P												P O P P P P																										
	T E R R R R						P O S S S S						T E R R R R						P O S S S S																				
	G A E E E E						T T T T T T						G A E E E E						T T T T T T																				
	R C M L S R						S M L S R S						R C M L S R						S M L S R S																				
	A H A A C E												A J A A C E												O A A C E O														
D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E	D E R H G E												
1	6	2	8	4	5	6	2	8	7	1	4	2	2	4	1	7	1	4	1	4	6	0	5	4	3	2	5	1	8	7	1	5	3	8	4	1	4	3	
1	6	2	7	2	4	0	8	8	0	3	2	3	2	5	1	7	3	1	8	3	7	3	2	7	1	2	4	1	3	8	0	2	5	2	8	0	4	6	4
2	6	2	2	7	4	1	4	3	5	5	4	1	2	4	1	3	8	4	7	1	6	6	5	8	1	2	5	1	8	0	2	5	2	8	0	4	6	4	
2	6	2	4	3	8	3	4	5	6	6	2	1	2	5	1	8	0	2	5	2	8	0	4	6	4	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	7	8	3	2	5	8	6	2	2	2	1	5	1	5	5	2	7	3	5	4	1	6	0	2	5	1	8	4	1	4	1	8	6	5	3	4	
1	6	2	5	5	6	4	2	6	6	4	1	3	1	4	1	7	2	7	0	3	7	5	6	0	4	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	2	8	7	6	1	2	8	6	7	4	1	5	1	6	3	1	5	0	7	2	0	3	0	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	0	1	5	4	7	1	0	3	6	7	2	4	1	6	2	5	6	5	8	3	0	6	1	2	5	1	8	0	2	5	2	8	0	4	6	4	
2	6	2	7	5	2	4	6	3	5	3	3	6	1	5	1	5	1	7	3	4	4	3	6	7	1	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	7	6	2	5	3	6	2	2	3	1	2	5	1	8	3	3	4	0	7	4	2	2	1	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	6	2	7	0	6	5	3	8	0	5	4	1	2	5	1	7	7	3	3	2	6	8	1	4	3	2	5	1	8	0	2	5	2	8	0	4	6	4	
1	6	2	7	0	5	2	3	5	5	4	1	0	1	5	1	4	6	8	2	5	6	4	7	3	8	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	2	5	6	7	8	4	3	1	8	5	2	5	1	7	4	0	1	2	7	2	2	4	2	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	6	2	5	4	6	1	8	1	2	5	3	6	2	4	1	6	4	2	5	8	7	1	3	4	2	4	1	3	8	5	4	1	6	0	5	3	4		
1	6	2	7	1	5	4	7	3	5	6	1	8	2	4	1	7	2	5	3	4	2	8	5	4	1	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	6	2	3	2	3	6	3	6	1	3	2	4	1	7	2	5	3	4	2	8	5	4	1	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	6	2	3	2	3	6	3	6	1	3	2	4	1	7	2	5	3	4	2	8	5	4	1	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	6	2	6	7	2	5	1	5	6	2	3	3	2	4	1	3	5	7	3	3	6	4	4	1	2	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	7	2	0	6	1	7	4	1	6	0	2	4	1	6	3	2	7	1	3	3	3	7	0	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	4	5	1	7	2	8	5	4	6	3	2	4	1	5	1	3	6	3	6	3	1	4	0	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	7	1	1	3	6	8	1	3	2	5	2	4	1	5	1	5	5	2	7	3	8	2	5	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	6	2	5	0	2	3	6	3	1	6	2	6	2	4	1	5	1	4	6	0	4	3	4	1	1	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	6	4	1	1	5	8	5	6	1	2	2	4	1	5	2	1	7	1	5	2	4	8	0	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	6	2	5	3	5	2	7	6	5	4	1	2	2	4	1	7	3	1	4	0	6	2	5	2	3	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	6	2	6	3	6	3	3	7	3	7	0	4	2	4	1	0	4	2	6	4	0	4	2	6	5	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	6	2	7	1	2	8	0	7	1	3	5	3	2	4	1	7	6	1	5	0	8	4	0	6	1	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	6	2	8	1	5	6	3	8	4	6	3	5	2	4	1	7	0	6	3	2	8	1	6	0	4	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	4	1	7	2	0	6	1	8	3	2	5	3	2	4	1	7	1	4	5	2	8	0	6	2	5	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	4	1	8	0	2	4	1	8	3	1	6	0	2	4	1	7	0	5	6	1	5	2	1	6	0	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	4	1	8	0	7	1	5	8	1	2	3	0	2	4	1	6	3	6	8	3	3	2	1	6	6	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	4	1	7	3	2	6	2	7	4	1	7	3	2	4	1	8	5	5	6	4	8	4	7	6	5	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	4	1	8	7	1	6	0	7	3	2	6	1	2	4	1	7	6	1	3	3	8	7	1	5	3	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	4	1	7	7	1	5	0	6	7	1	0	4	2	4	1	7	1	5	6	3	7	4	1	5	0	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	4	1	6	6	3	7	1	8	5	4	3	2	2	4	1	8	5	3	5	2	8	5	1	6	0	2	4	1	3	8	5	4	1	6	0	5	3	4	
1	4	1	4	0	4	3	6	2	0	7	1	6	2	4	1	3	6	5	1	2	6	4	3	0	2	2	4	1	3	8	5	4	1	6	0	5	3	4	
2	4	1	6	3	1	7	0	0	3	4	7	4	2	4	1	7	4	0	4	2	7	4	1	6	0	2	4	1	3	8	5	4	1	6	0	5	3	4	

(Continued)

ATTITUDE DATA (Continued)

P P P P P										P P P P P														
T P P P P P O O O O O										T P P P P P O O O O O														
E R R R R R S S S S S										E R R R R R S S S S S														
G A E E E E T T T T T										G A E E E E T T T T T														
R C M L S R S M L S R S										R C M L S R S M L S R S														
S A H A A C E O A A C E O										S A J A A C E O A A C E O														
E D E T N I A C T N I A C										E D H T N I A C T N I A C														
X E R H G E D S H G E D S										X E R H G E D S H G E D S														
2 4	1	7	4	2	6	3	7	5	1	6	0	1	5	1	4	0	6	7	6	1	0	5	8	7
1 5	1	6	4	4	5	4	7	2	5	4	5	2	5	1	7	4	5	1	4	8	5	2	4	2
2 5	1	5	5	0	6	3	5	5	0	6	1	1	5	1	7	4	3	8	2	6	2	3	7	4
1 5	1	6	1	3	1	1	6	5	3	1	0	2	5	1	7	5	2	6	2	7	6	0	6	2
2 5	1	8	4	2	6	3	8	2	4	6	6	2	5	1	6	3	3	7	5	4	4	0	7	6
1 5	1	6	2	5	3	3	8	3	4	1	5	2	5	1	4	2	5	5	4	7	4	3	7	0
2 5	1	7	4	2	5	0	7	4	1	7	2	2	5	1	6	3	4	6	3	5	4	7	6	5
2 5	1	7	4	0	6	1	8	3	0	7	0	1	4	1	6	1	4	8	1	1	2	3	7	1
2 4	1	4	1	7	1	5	4	2	8	2	7	2	5	1	8	4	2	4	3	8	4	0	3	3
1 4	1	6	0	7	5	4	5	1	4	7	5	2	5	1	7	5	0	8	3	7	6	1	8	5
2 5	1	7	4	3	6	5	7	3	1	6	4	1	4	1	7	1	5	6	0	7	5	5	4	3
2 4	1	6	0	6	1	5	8	2	4	3	4	1	4	1	5	3	6	0	3	8	3	6	0	5
2 4	1	8	5	2	4	0	8	5	3	5	0	2	4	1	8	2	0	7	4	4	4	0	8	5
2 4	1	6	7	3	3	6	7	2	4	6	2	2	5	1	4	4	2	6	1	7	3	4	7	5
2 4	1	7	6	2	4	0	8	4	1	5	0	1	5	1	7	2	7	3	1	7	4	5	0	1
1 4	1	3	4	2	4	0	6	3	3	3	3	2	4	1	7	4	2	6	5	6	3	2	5	4
1 4	1	8	4	5	0	3	8	4	0	5	2	2	6	2	3	2	0	2	3	3	3	0	2	3
2 4	1	8	0	4	1	4	8	6	4	1	3	2	4	1	6	4	4	7	2	6	5	5	4	1
2 4	1	7	6	5	2	3	7	5	2	3	1													

DESCRIPTION OF NATIONAL ACADEMY
OF SCIENCES DATA

The following data is from a 1982 National Academy of Sciences published report rating the “scholarly quality” of research programs in the humanities, physical sciences and social sciences. The ratings were based on the rankings of quality and reputation made by senior faculty in the field who taught at institutions other than the one being rated.

The data to be presented are the quality ratings of 46 research doctorate programs in psychology, as well as six potential correlates of the quality ratings. Here is a description of the variables: **QUALITY** Mean rating of scholarly quality of program faculty **NFACULTY** Number of faculty members in program as of December 1980 **NGRADS** Number of program graduates from 1975 through 1980 **PCTSUPP** Percentage of program graduates from 1975–1979 that received fellowships or training grant support during their graduate education **PCTGRANT** Percent of faculty members holding research grants from the Alcohol, Drug Abuse and Mental Health Administration, the National Institute of Health or the National Science Foundation at any time during 1978–1980 **NARTICLE** Number of published articles attributed to program faculty members 1978–1980 **PCTPUB** Percent of faculty with one or more published articles from 1978–1980.

O B S	NAME	Q	N		P	P	N	
		U	F		C	C	A	
		A	A	N	T	T	R	P
		L	C	G	S	G	T	C
		I	U	R	U	R	I	T
		T	L	A	P	A	C	P
		Y	Y	S	P	T	L	U
							E	B
1	ADELPHI	12	13	19	16	8	14	39
2	ARIZONA-TUSCON	23	29	72	67	3	61	66
3	BOSTON UNIV	29	38	111	66	13	68	68
4	BROWN	36	16	28	52	63	49	75
5	U C BERKELEY	44	40	104	64	53	130	83
6	U C RIVERSIDE	21	14	28	59	29	65	79
7	CARNEGIE MELLON	40	44	16	81	35	79	82
8	UNIV OF CHICAGO	42	60	57	65	40	187	82
9	CLARK UNIV	24	16	18	87	19	32	75
10	COLUMBIA TEACHERS	30	37	41	43	8	50	54
11	DELAWARE, UNIV OF	20	20	45	26	25	49	50
12	DETROIT, UNIV OF	8	11	27	7	0	9	27
13	FLORIDA ST-TALAH	28	29	112	64	35	65	69
14	FULLER THEOL SEMIN	14	14	57	10	0	11	43
15	UNIV OF GEORGIA	27	38	167	28	13	196	84
16	HARVARD	46	27	113	62	52	173	85
17	HOUSTON, UNIV OF	29	32	122	51	119	79	69
18	UNIV ILLINOIS-CHAMP	42	56	116	56	32	208	73
19	IOWA, UNIV OF	33	32	54	49	10	120	69
20	KANSAS, UNIV OF	31	42	79	41	14	14	71
21	KENT STATE UNIV	23	30	76	22	20	87	67
22	LOUISIANA STATE	18	18	62	39	6	10	39
23	UNIV OF MARYLAND	29	41	98	41	12	101	66
24	MIAMI UNIV	21	23	52	33	4	59	78
25	U MICH-ANN ARB	45	111	222	64	32	274	70
26	U MISSOURI	25	26	63	39	23	160	89
27	U NEW HAMPSHIRE	18	16	24	4	31	39	63
28	NEW YORK UNIV	33	38	154	55	34	84	63
29	U NC—GREENSBORO	21	19	40	7	5	60	84
30	NORTHEASTERN	24	16	18	25	63	31	63
31	NOTRE DAME	15	13	29	23	15	62	85
32	OKLA ST-STILLWATER	15	23	41	51	4	24	57
33	PENN STATE	36	32	69	65	16	122	75
34	PRINCETON	38	21	38	28	48	92	91
35	UNIV OF ROCHESTER	32	28	90	70	36	117	61
36	SUNY ALBANY	27	22	52	10	27	114	86
37	ST LOUIS UNIVERSITY	16	20	80	46	10	19	40
38	UNIV SOUTH FLORIDA	26	32	41	13	6	64	56
39	STANFORD	48	26	81	70	58	155	100
40	TEMPLE	26	40	81	42	10	70	68

(Continued)

O B S	NAME	Q	N		P	P	N	
		U	F		C	C	A	P
		A	A	N	C	T	R	C
		L	C	G	T	G	T	T
		I	U	R	S	R	I	P
		T	L	A	U	A	C	P
		T	T	D	P	N	L	U
		Y	Y	S	P	T	E	B
41	TEXAS TECH LUBBOCK	14	19	87	15	5	72	79
42	UNIV OF TOLEDO	12	17	26	9	6	15	59
43	UNIV OF UTAH, SALT L	29	29	71	74	17	85	76
44	VIRGINIA POLYTECH	34	27	20	0	29	79	57
45	WASHINGTON UNIV-ST. L	28	26	70	68	27	84	73
46	UNIV WISC—MADISON	39	36	59	57	67	172	83

Jones, L. V., Lindzey, G., & Coggeshall, P. (Eds.) (1982). *An Assessment of Research-Doctorate Programs in the United States: Social and Behavioral Sciences*. (Washington, DC: National Academy Press).

AGRESTI HOME SALES DATA

	<i>price</i>	<i>size</i>	<i>nobed</i>	<i>nobath</i>	<i>new</i>
1	48.50	1.10	3.00	1.00	.00
2	55.00	1.01	3.00	2.00	.00
3	68.00	1.45	3.00	2.00	.00
4	137.00	2.40	3.00	3.00	.00
5	309.40	3.30	4.00	3.00	1.00
6	17.50	.40	1.00	1.00	.00
7	19.60	1.28	3.00	1.00	.00
8	24.50	.74	3.00	1.00	.00
9	34.80	.78	2.00	1.00	.00
10	32.00	.97	3.00	1.00	.00
11	28.00	.84	3.00	1.00	.00
12	49.90	1.08	2.00	2.00	.00
13	59.90	.99	2.00	1.00	.00
14	61.50	1.01	3.00	2.00	.00
15	60.00	1.34	3.00	2.00	.00
16	65.90	1.22	3.00	1.00	.00
17	67.90	1.28	3.00	2.00	.00
18	68.90	1.29	3.00	2.00	.00
19	69.90	1.52	3.00	2.00	.00
20	70.50	1.25	3.00	2.00	.00
21	72.90	1.28	3.00	2.00	.00
22	72.50	1.28	3.00	1.00	.00
23	72.00	1.36	3.00	2.00	.00
24	71.00	1.20	3.00	2.00	.00
25	76.00	1.46	3.00	2.00	.00
26	72.90	1.56	4.00	2.00	.00
27	73.00	1.22	3.00	2.00	.00
28	70.00	1.40	2.00	2.00	.00
29	76.00	1.15	2.00	2.00	.00
30	69.00	1.74	3.00	2.00	.00
31	75.50	1.62	3.00	2.00	.00
32	76.00	1.66	3.00	2.00	.00
33	81.80	1.33	3.00	2.00	.00
34	84.50	1.34	3.00	2.00	.00
35	83.50	1.40	3.00	2.00	.00
36	86.00	1.15	2.00	2.00	1.00
37	86.90	1.58	3.00	2.00	1.00
38	86.90	1.58	3.00	2.00	1.00
39	86.90	1.58	3.00	2.00	1.00
40	87.90	1.71	3.00	2.00	.00
41	88.10	2.10	3.00	2.00	.00
42	85.90	1.27	3.00	2.00	.00
43	89.50	1.34	3.00	2.00	.00
44	87.40	1.25	3.00	2.00	.00
45	87.90	1.68	3.00	2.00	.00

(Continued)

AGRESTI HOME SALES DATA (Continued)

	<i>price</i>	<i>size</i>	<i>nobed</i>	<i>nobath</i>	<i>new</i>
46	88.00	1.55	3.00	2.00	.00
47	90.00	1.55	3.00	2.00	.00
48	96.00	1.36	3.00	2.00	1.00
49	99.90	1.51	3.00	2.00	1.00
50	95.50	1.54	3.00	2.00	1.00
51	98.50	1.51	3.00	2.00	.00
52	100.10	1.85	3.00	2.00	.00
53	99.90	1.62	4.00	2.00	1.00
54	101.90	1.40	3.00	2.00	1.00
55	101.90	1.92	4.00	2.00	.00
56	102.30	1.42	3.00	2.00	1.00
57	110.80	1.56	3.00	2.00	1.00
58	105.00	1.43	3.00	2.00	1.00
59	97.90	2.00	3.00	2.00	.00
60	106.30	1.45	3.00	2.00	1.00
61	106.50	1.65	3.00	2.00	.00
62	116.00	1.72	4.00	2.00	1.00
63	108.00	1.79	4.00	2.00	1.00
64	107.50	1.85	3.00	2.00	.00
65	109.90	2.06	4.00	2.00	1.00
66	110.00	1.76	4.00	2.00	.00
67	120.00	1.62	3.00	2.00	1.00
68	115.00	1.80	4.00	2.00	1.00
69	113.40	1.98	3.00	2.00	.00
70	114.90	1.57	3.00	2.00	.00
71	115.00	2.19	3.00	2.00	.00
72	115.00	2.07	4.00	2.00	.00
73	117.90	1.99	4.00	2.00	.00
74	110.00	1.55	3.00	2.00	.00
75	115.00	1.67	3.00	2.00	.00
76	124.00	2.40	4.00	2.00	.00
77	129.90	1.79	4.00	2.00	1.00
78	124.00	1.89	3.00	2.00	.00
79	128.00	1.88	3.00	2.00	1.00
80	132.40	2.00	4.00	2.00	1.00
81	139.30	2.05	4.00	2.00	1.00
82	139.30	2.00	4.00	2.00	1.00
83	139.70	2.03	3.00	2.00	1.00
84	142.00	2.12	3.00	3.00	.00
85	141.30	2.08	4.00	2.00	1.00

(Continued)

AGRESTI HOME SALES DATA (Continued)

	<i>price</i>	<i>size</i>	<i>nobed</i>	<i>nobath</i>	<i>new</i>
86	147.50	2.19	4.00	2.00	.00
87	142.50	2.40	4.00	2.00	.00
88	148.00	2.40	5.00	2.00	.00
89	149.00	3.05	4.00	2.00	.00
90	150.00	2.04	3.00	3.00	.00
91	172.90	2.25	4.00	2.00	1.00
92	190.00	2.57	4.00	3.00	1.00
93	280.00	3.85	4.00	3.00	.00

Appendix B

Statistical Tables

CONTENTS

- Table B.1 Critical Values for F
- Table B.2 Percentile Points of Studentized Range Statistic
- Table B.3 Critical Values for Dunnett's Test
- Table B.4 Critical Values for F (max) Statistic
- Table B.5 Critical Values for Bryant-Paulson Procedure

TABLE B.1
Critical Values for *F*

		<i>df for Numerator</i>							
<i>df error</i>	α	1	2	3	4	5	6	8	12
1	.01	4052	4999	5403	5625	5764	5859	5981	6106
	.05	161.45	199.50	215.71	224.58	230.16	233.99	238.88	243.91
	.10	39.85	49.50	53.59	55.83	57.24	58.20	59.44	60.70
	.20	9.47	12.00	13.06	13.73	14.01	14.26	14.59	14.90
2	.01	98.49	99.00	99.17	99.25	99.30	99.33	99.36	99.42
	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41
	.10	8.53	9.00	9.16	9.24	9.29	9.33	9.37	9.41
	.20	3.56	4.00	4.16	4.24	4.28	4.32	4.36	4.40
3	.001	167.5	148.5	141.1	137.1	134.6	132.8	130.6	128.3
	.01	34.12	30.81	29.46	28.71	28.24	27.91	27.49	27.05
	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74
	.10	5.54	5.46	5.39	5.34	5.31	5.28	5.25	5.22
	.20	2.68	2.89	2.94	2.96	2.97	2.97	2.98	2.98
4	.001	74.14	61.25	56.18	53.44	51.71	50.53	49.00	47.41
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.80	14.37
	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91
	.10	4.54	4.32	4.19	4.11	4.05	4.01	3.95	3.90
	.20	2.35	2.47	2.48	2.48	2.48	2.47	2.47	2.46
5	.001	47.04	36.61	33.20	31.09	29.75	28.84	27.64	26.42
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.29	9.89
	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.68
	.10	4.06	3.78	3.62	3.52	3.45	3.40	3.34	3.27
	.20	2.18	2.26	2.25	2.24	2.23	2.22	2.20	2.18
6	.001	35.51	27.00	23.70	21.90	20.81	20.03	19.03	17.99
	.01	13.74	10.92	9.78	9.15	8.75	8.47	8.10	7.72
	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.15	4.00
	.10	3.78	3.46	3.29	3.18	3.11	3.05	2.98	2.90
	.20	2.07	2.13	2.11	2.09	2.08	2.06	2.04	2.02
7	.001	29.22	21.69	18.77	17.19	16.21	15.52	14.63	13.71
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.84	6.47
	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57
	.10	3.59	3.26	3.07	2.96	2.88	2.83	2.75	2.67
	.20	2.00	2.04	2.02	1.99	1.97	1.96	1.93	1.91

(Continued)

TABLE B.1 (Continued)

<i>df error</i>	α	<i>df for Numerator</i>							
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>8</i>	<i>12</i>
8	.001	25.42	18.49	15.83	14.39	13.49	12.86	12.04	11.19
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.03	5.67
	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.44	3.28
	.10	3.46	3.11	2.92	2.81	2.73	2.67	2.59	2.50
	.20	1.95	1.98	1.95	1.92	1.90	1.88	1.86	1.83
9	.001	22.86	16.39	13.90	12.56	11.71	11.13	10.37	9.57
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.47	5.11
	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.23	3.07
	.10	3.36	3.01	2.81	2.69	2.61	2.55	2.47	2.38
	.20	1.91	1.94	1.90	1.87	1.85	1.83	1.80	1.76
10	.001	21.04	14.91	12.55	11.28	10.48	9.92	9.20	8.45
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.06	4.71
	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.07	2.91
	.10	3.28	2.92	2.73	2.61	2.52	2.46	2.38	2.28
	.20	1.88	1.90	1.86	1.83	1.80	1.78	1.75	1.72
11	.001	19.69	13.81	11.56	10.35	9.58	9.05	8.35	7.63
	.01	9.65	7.20	6.22	5.67	5.32	5.07	4.74	4.40
	.05	4.84	3.98	3.59	3.36	3.20	3.09	2.95	2.79
	.10	3.23	2.86	2.66	2.54	2.45	2.39	2.30	2.21
	.20	1.86	1.87	1.83	1.80	1.77	1.75	1.72	1.68
12	.001	18.64	12.97	10.80	9.63	8.89	8.38	7.71	7.00
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.50	4.16
	.05	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69
	.10	3.18	2.81	2.61	2.48	2.39	2.33	2.24	2.15
	.20	1.84	1.85	1.80	1.77	1.74	1.72	1.69	1.65
13	.001	17.81	12.31	10.21	9.07	8.35	7.86	7.21	6.52
	.01	9.07	6.70	5.74	5.20	4.86	4.62	4.30	3.96
	.05	4.67	3.80	3.41	3.18	3.02	2.92	2.77	2.60
	.10	3.14	2.76	2.56	2.43	2.35	2.28	2.20	2.10
	.20	1.82	1.83	1.78	1.75	1.72	1.69	1.66	1.62
14	.001	17.14	11.78	9.73	8.62	7.92	7.43	6.80	6.13
	.01	8.86	6.51	5.56	5.03	4.69	4.46	4.14	3.80
	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.70	2.53
	.10	3.10	2.73	2.52	2.39	2.31	2.24	2.15	2.05
	.20	1.81	1.81	1.76	1.73	1.70	1.67	1.64	1.60

(Continued)

TABLE B.1 (Continued)

df error	α	df for Numerator							
		1	2	3	4	5	6	8	12
15	.001	16.59	11.34	9.34	8.25	7.57	7.09	6.47	5.81
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.00	3.67
	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.48
	.10	3.07	2.70	2.49	2.36	2.27	2.21	2.12	2.02
	.20	1.80	1.79	1.75	1.71	1.68	1.66	1.62	1.58
16	.001	16.12	10.97	9.00	7.94	7.27	6.81	6.19	5.55
	.01	8.53	6.23	5.29	4.77	4.44	4.20	3.89	3.55
	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.42
	.10	3.05	2.67	2.46	2.33	2.24	2.18	2.09	1.99
	.20	1.79	1.78	1.74	1.70	1.67	1.64	1.61	1.56
17	.001	15.72	10.66	8.73	7.68	7.02	6.56	5.96	5.32
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.79	3.45
	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.55	2.38
	.10	3.03	2.64	2.44	2.31	2.22	2.15	2.06	1.96
	.20	1.78	1.77	1.72	1.68	1.65	1.63	1.59	1.55
18	.001	15.38	10.39	8.49	7.46	6.81	6.35	5.76	5.13
	.01	8.28	6.01	5.09	4.58	4.25	4.01	3.71	3.37
	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.51	2.34
	.10	3.01	2.62	2.42	2.29	2.20	2.13	2.04	1.93
	.20	1.77	1.76	1.71	1.67	1.64	1.62	1.58	1.53
19	.001	15.08	10.16	8.28	7.26	6.61	6.18	5.59	4.97
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.63	3.30
	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.48	2.31
	.10	2.99	2.61	2.40	2.27	2.18	2.11	2.02	1.91
	.20	1.76	1.75	1.70	1.66	1.63	1.61	1.57	1.52
20	.001	14.82	9.95	8.10	7.10	6.46	6.02	5.44	4.82
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.56	3.23
	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.28
	.10	2.97	2.59	2.38	2.25	2.16	2.09	2.00	1.89
	.20	1.76	1.75	1.70	1.65	1.62	1.60	1.56	1.51
21	.001	14.59	9.77	7.94	6.95	6.32	5.88	5.31	4.70
	.01	8.02	5.78	4.87	4.37	4.04	3.81	3.51	3.17
	.05	4.32	3.47	3.07	2.84	2.68	2.57	2.42	2.25
	.10	2.96	2.57	2.36	2.23	2.14	2.08	1.98	1.88
	.20	1.75	1.74	1.69	1.65	1.61	1.59	1.55	1.50

(Continued)

TABLE B.1 (Continued)

df error	α	df for Numerator							
		1	2	3	4	5	6	8	12
22	.001	14.38	9.61	7.80	6.81	6.19	5.76	5.19	4.58
	.01	7.94	5.72	4.82	4.31	3.99	3.76	3.45	3.12
	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.40	2.23
	.10	2.95	2.56	2.35	2.22	2.13	2.06	1.97	1.86
	.20	1.75	1.73	1.68	1.64	1.61	1.58	1.54	1.49
23	.001	14.19	9.47	7.67	6.69	6.08	5.65	5.09	4.48
	.01	7.88	5.66	4.76	4.26	3.94	3.71	3.41	3.07
	.05	4.28	3.42	3.03	2.80	2.64	2.53	2.38	2.20
	.10	2.94	2.55	2.34	2.21	2.11	2.05	1.95	1.84
	.20	1.74	1.73	1.68	1.63	1.60	1.57	1.53	1.49
24	.001	14.03	9.34	7.55	6.59	5.98	5.55	4.99	4.39
	.01	7.82	5.61	4.72	4.22	3.90	3.67	3.36	3.03
	.05	4.26	3.40	3.01	2.78	2.62	2.51	2.36	2.18
	.10	2.93	2.54	2.33	2.19	2.10	2.04	1.94	1.83
	.20	1.74	1.72	1.67	1.63	1.59	1.57	1.53	1.48
25	.001	13.88	9.22	7.45	6.49	5.88	5.46	4.91	4.31
	.01	7.77	5.57	4.68	4.18	3.86	3.63	3.32	2.99
	.05	4.24	3.38	2.99	2.76	2.60	2.49	2.34	2.16
	.10	2.92	2.53	2.32	2.18	2.09	2.02	1.93	1.82
	.20	1.73	1.72	1.66	1.62	1.59	1.56	1.52	1.47
26	.001	13.74	9.12	7.36	6.41	5.80	5.38	4.83	4.24
	.01	7.72	5.53	4.64	4.14	3.82	3.59	3.29	2.96
	.05	4.22	3.37	2.98	2.74	2.59	2.47	2.32	2.15
	.10	2.91	2.52	2.31	2.17	2.08	2.01	1.92	1.81
	.20	1.73	1.71	1.66	1.62	1.58	1.56	1.52	1.47
27	.001	13.61	9.02	7.27	6.33	5.73	5.31	4.76	4.17
	.01	7.68	5.49	4.60	4.11	3.78	3.56	3.26	2.93
	.05	4.21	3.35	2.96	2.73	2.57	2.46	2.30	2.13
	.10	2.90	2.51	2.30	2.17	2.07	2.00	1.91	1.80
	.20	1.73	1.71	1.66	1.61	1.58	1.55	1.51	1.46
28	.001	13.50	8.93	7.19	6.25	5.66	5.24	4.69	4.11
	.01	7.64	5.45	4.57	4.07	3.75	3.53	3.23	2.90
	.05	4.20	3.34	2.95	2.71	2.56	2.44	2.29	2.12
	.10	2.89	2.50	2.29	2.16	2.06	2.00	1.90	1.79
	.20	1.72	1.71	1.65	1.61	1.57	1.55	1.51	1.46

(Continued)

TABLE B.1 (Continued)

df error	α	df for Numerator							
		1	2	3	4	5	6	8	12
29	.001	13.39	8.85	7.12	6.19	5.59	5.18	4.64	4.05
	.01	7.60	5.42	4.54	4.04	3.73	3.50	3.20	2.87
	.05	4.18	3.33	2.93	2.70	2.54	2.43	2.28	2.10
	.10	2.89	2.50	2.28	2.15	2.06	1.99	1.89	1.78
	.20	1.72	1.70	1.65	1.60	1.57	1.54	1.50	1.45
30	.001	13.29	8.77	7.05	6.12	5.53	5.12	4.58	4.00
	.01	7.56	5.39	4.51	4.02	3.70	3.47	3.17	2.84
	.05	4.17	3.32	2.92	2.69	2.53	2.42	2.27	2.09
	.10	2.88	2.49	2.28	2.14	2.05	1.98	1.88	1.77
	.20	1.72	1.70	1.64	1.60	1.57	1.54	1.50	1.45
40	.001	12.61	8.25	6.60	5.70	5.13	4.73	4.21	3.64
	.01	7.31	5.18	4.31	3.83	3.51	3.29	2.99	2.66
	.05	4.08	3.23	2.84	2.61	2.45	2.34	2.18	2.00
	.10	2.84	2.44	2.23	2.09	2.00	1.93	1.83	1.71
	.20	1.70	1.68	1.62	1.57	1.54	1.51	1.47	1.41
60	.001	11.97	7.76	6.17	5.31	4.76	4.37	3.87	3.31
	.01	7.08	4.98	4.13	3.65	3.34	3.12	2.82	2.50
	.05	4.00	3.15	2.76	2.52	2.37	2.25	2.10	1.92
	.10	2.79	2.39	2.18	2.04	1.95	1.87	1.77	1.66
	.20	1.68	1.65	1.59	1.55	1.51	1.48	1.44	1.38
120	.001	11.38	7.31	5.79	4.95	4.42	4.04	3.55	3.02
	.01	6.85	4.79	3.95	3.48	3.17	2.96	2.66	2.34
	.05	3.92	3.07	2.68	2.45	2.29	2.17	2.02	1.83
	.10	2.75	2.35	2.13	1.99	1.90	1.82	1.72	1.60
	.20	1.66	1.63	1.57	1.52	1.48	1.45	1.41	1.35
∞	.001	10.83	6.91	5.42	4.62	4.10	3.74	3.27	2.74
	.01	6.64	4.60	3.78	3.32	3.02	2.80	2.51	2.18
	.05	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75
	.10	2.71	2.30	2.08	1.94	1.85	1.77	1.67	1.55
	.20	1.64	1.61	1.55	1.50	1.46	1.43	1.38	1.32

Source: Reproduced from E. F. Lindquist, *Design and Analysis of Experiments in Psychology and Education*, Houghton Mifflin, Boston, 1953, pp. 41–44, with the permission of the publisher.

TABLE B.2
Percentile Points of Studentized Range Statistic

<i>90th Percentile</i>									
<i>number of groups</i>									
<i>df error</i>	2	3	4	5	6	7	8	9	10
1	8.929	13.44	16.36	18.49	20.15	21.51	22.64	23.62	24.48
2	4.130	5.733	6.773	7.538	8.139	8.633	9.049	9.409	9.725
3	3.328	4.467	5.199	5.738	6.162	6.511	6.806	7.062	7.287
4	3.015	3.976	4.586	5.035	5.388	5.679	5.926	6.139	6.327
5	2.850	3.717	4.264	4.664	4.979	5.238	5.458	5.648	5.816
6	2.748	3.559	4.065	4.435	4.726	4.966	5.168	5.344	5.499
7	2.680	3.451	3.931	4.280	4.555	4.780	4.972	5.137	5.283
8	2.630	3.374	3.834	4.169	4.431	4.646	4.829	4.987	5.126
9	2.592	3.316	3.761	4.084	4.337	4.545	4.721	4.873	5.007
10	2.563	3.270	3.704	4.018	4.264	4.465	4.636	4.783	4.913
11	2.540	3.234	3.658	3.965	4.205	4.401	4.568	4.711	4.838
12	2.521	3.204	3.621	3.922	4.156	4.349	4.511	4.652	4.776
13	2.505	3.179	3.589	3.885	4.116	4.305	4.464	4.602	4.724
14	2.491	3.158	3.563	3.854	4.081	4.267	4.424	4.560	4.680
15	2.479	3.140	3.540	3.828	4.052	4.235	4.390	4.524	4.641
16	2.469	3.124	3.520	3.804	4.026	4.207	4.360	4.492	4.608
17	2.460	3.110	3.503	3.784	4.004	4.183	4.334	4.464	4.579
18	2.452	3.098	3.488	3.767	3.984	4.161	4.311	4.440	4.554
19	2.445	3.087	3.474	3.751	3.966	4.142	4.290	4.418	4.531
20	2.439	3.078	3.462	3.736	3.950	4.124	4.271	4.398	4.510
24	2.420	3.047	3.423	3.692	3.900	4.070	4.213	4.336	4.445
30	2.400	3.017	3.386	3.648	3.851	4.016	4.155	4.275	4.381
40	2.381	2.988	3.349	3.605	3.803	3.963	4.099	4.215	4.317
60	2.363	2.959	3.312	3.562	3.755	3.911	4.042	4.155	4.254
120	2.344	2.930	3.276	3.520	3.707	3.859	3.987	4.096	4.191
∞	2.326	2.902	3.240	3.478	3.661	3.808	3.931	4.037	4.129

(Continued)

TABLE 3.2 (Continued)

<i>95th Percentiles</i>									
<i>number of groups</i>									
<i>df error</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	<i>7</i>	<i>8</i>	<i>9</i>	<i>10</i>
1	17.97	26.98	32.82	37.08	40.41	43.12	45.40	47.36	49.07
2	6.085	8.331	9.798	10.88	11.74	12.44	13.03	13.54	13.99
3	4.501	5.910	6.825	7.502	8.037	8.478	8.853	9.177	9.462
4	3.927	5.040	5.757	6.287	6.707	7.053	7.347	7.602	7.826
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.802	6.995
6	3.461	4.339	4.896	5.305	5.628	5.895	6.122	6.319	6.493
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.998	6.158
8	3.261	4.041	4.529	4.886	5.167	5.399	5.597	5.767	5.918
9	3.199	3.949	4.415	4.756	5.024	5.244	5.432	5.595	5.739
10	3.151	3.877	4.327	4.654	4.912	5.124	5.305	5.461	5.599
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.487
12	3.082	3.773	4.199	4.508	4.751	4.950	5.119	5.265	5.395
13	3.055	3.735	4.151	4.453	4.690	4.885	5.049	5.192	5.318
14	3.033	3.702	4.111	4.407	4.639	4.829	4.990	5.131	5.254
15	3.014	3.674	4.076	4.367	4.595	4.782	4.940	5.077	5.198
16	2.998	3.649	4.046	4.333	4.557	4.741	4.897	5.031	5.150
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108
18	2.971	3.609	3.997	4.277	4.495	4.673	4.824	4.956	5.071
19	2.960	3.593	3.977	4.253	4.469	4.645	4.794	4.924	5.038
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.896	5.008
24	2.919	3.532	3.901	4.166	4.373	4.541	4.684	4.807	4.915
30	2.888	3.486	3.845	4.102	4.302	4.464	4.602	4.720	4.824
40	2.858	3.442	3.791	4.039	4.232	4.389	4.521	4.635	4.735
60	2.829	3.399	3.737	3.977	4.163	4.314	4.441	4.550	4.646
120	2.800	3.356	3.685	3.917	4.096	4.241	4.363	4.468	4.560
∞	2.772	3.314	3.633	3.858	4.030	4.170	4.286	4.387	4.474

TABLE B.3
Critical Values for Dunnett's Test

<i>Two-Tailed Comparisons</i>										
<i>k = number of treatment means, including control</i>										
<i>df Error</i>	α	2	3	4	5	6	7	8	9	10
5	0.05	2.57	3.03	3.29	3.48	3.62	3.73	3.82	3.90	3.97
	0.01	4.03	4.63	4.98	5.22	5.41	5.56	5.69	5.80	5.89
6	0.05	2.45	2.86	3.10	3.26	3.39	3.49	3.57	3.64	3.71
	0.01	3.71	4.21	4.51	4.71	4.87	5.00	5.10	5.20	5.28
7	0.05	2.36	2.75	2.97	3.12	3.24	3.33	3.41	3.47	3.53
	0.01	3.50	3.95	4.21	4.39	4.53	4.64	4.74	4.82	4.89
8	0.05	2.31	2.67	2.88	3.02	3.13	3.22	3.29	3.35	3.41
	0.01	3.36	3.77	4.00	4.17	4.29	4.40	4.48	4.56	4.62
9	0.05	2.26	2.61	2.81	2.95	3.05	3.14	3.20	3.26	3.32
	0.01	3.25	3.63	3.85	4.01	4.12	4.22	4.30	4.37	4.43
10	0.05	2.23	2.57	2.76	2.89	2.99	3.07	3.14	3.19	3.24
	0.01	3.17	3.53	3.74	3.88	3.99	4.08	4.16	4.22	4.28
11	0.05	2.20	2.53	2.72	2.84	2.94	3.02	3.08	3.14	3.19
	0.01	3.11	3.45	3.65	3.79	3.89	3.98	4.05	4.11	4.16
12	0.05	2.18	2.50	2.68	2.81	2.90	2.98	3.04	3.09	3.14
	0.01	3.05	3.39	3.58	3.71	3.81	3.89	3.96	4.02	4.07
13	0.05	2.16	2.48	2.65	2.78	2.87	2.94	3.00	3.06	3.10
	0.01	3.01	3.33	3.52	3.65	3.74	3.82	3.89	3.94	3.99
14	0.05	2.14	2.46	2.63	2.75	2.84	2.91	2.97	3.02	3.07
	0.01	2.98	3.29	3.47	3.59	3.69	3.76	3.83	3.88	3.93
15	0.05	2.13	2.44	2.61	2.73	2.82	2.89	2.95	3.00	3.04
	0.01	2.95	3.25	3.43	3.55	3.64	3.71	3.78	3.83	3.88
16	0.05	2.12	2.42	2.59	2.71	2.80	2.87	2.92	2.97	3.02
	0.01	2.92	3.22	3.39	3.51	3.60	3.67	3.73	3.78	3.83
17	0.05	2.11	2.41	2.58	2.69	2.78	2.85	2.90	2.95	3.00
	0.01	2.90	3.19	3.36	3.47	3.56	3.63	3.69	3.74	3.79
18	0.05	2.10	2.40	2.56	2.68	2.76	2.83	2.89	2.94	2.98
	0.01	2.88	3.17	3.33	3.44	3.53	3.60	3.66	3.71	3.75
19	0.05	2.09	2.39	2.55	2.66	2.75	2.81	2.87	2.92	2.96
	0.01	2.86	3.15	3.31	3.42	3.50	3.57	3.63	3.68	3.72
20	0.05	2.09	2.38	2.54	2.65	2.72	2.80	2.86	2.90	2.95
	0.01	2.85	3.13	3.29	3.40	3.48	3.55	3.60	3.65	3.69
24	0.05	2.06	2.35	2.51	2.61	2.70	2.76	2.81	2.86	2.90
	0.01	2.80	3.07	3.22	3.32	3.40	3.47	3.52	3.57	3.61
30	0.05	2.04	2.32	2.47	2.58	2.66	2.72	2.77	2.82	2.86
	0.01	2.75	3.01	3.15	3.25	3.33	3.39	3.44	3.49	3.52
40	0.05	2.02	2.29	2.44	2.54	2.62	2.68	2.73	2.77	2.81
	0.01	2.70	2.95	3.09	3.19	3.26	3.32	3.37	3.41	3.44
60	0.05	2.00	2.27	2.41	2.51	2.58	2.64	2.69	2.73	2.77
	0.01	2.66	2.90	3.03	3.12	3.19	3.25	3.29	3.33	3.37
120	0.05	1.98	2.24	2.38	2.47	2.55	2.60	2.65	2.69	2.73
	0.01	2.62	2.85	2.97	3.06	3.12	3.18	3.22	3.26	3.29
∞	0.05	1.96	2.21	2.35	2.44	2.51	2.57	2.61	2.65	2.69
	0.01	2.58	2.79	2.92	3.00	3.06	3.11	3.15	3.19	3.22

Reproduced from: C. W. Dunnett (1964). New tables for multiple comparisons with a control, *Biometrics* 20, 482–491. With permission of The Biometric Society.

TABLE B.4
Critical Values for F_{\max} Statistic

df for * each variance	1-α	Number of Variances each										
		2	3	4	5	6	7	8	9	10	11	12
2	.95	39.0	87.5	142	202	266	333	403	475	550	626	704
	.99	199	448	729	1036	1362	1705	2063	2432	2813	3204	3605
3	.95	154	27.8	39.2	50.7	62.0	72.9	83.5	93.9	104	114	124
	.99	47.5	85	120	151	184	216	249	281	310	337	361
4	.95	9.60	15.5	20.6	25.2	29.5	33.6	37.5	41.4	44.6	48.0	51.4
	.99	23.2	37.	49.	59.	69.	79.	89.	97.	106.	113.	120
5	.95	7.15	10.8	13.7	16.3	18.7	20.8	22.9	24.7	26.5	28.2	29.9
	.99	14.9	22.	28.	33.	38.	42.	46.	50.	54.	57	60
6	.95	5.82	8.38	10.4	12.1	13.7	15.0	16.3	17.5	18.6	19.7	20.7
	.99	11.1	15.5	19.1	22.	25.	27.	30.	32.	34.	36	37
7	.95	4.99	6.94	8.44	9.70	10.8	11.8	12.7	13.5	14.3	151	15.8
	.99	8.89	12.1	14.5	16.5	18.4	20.	22.	23.	24.	26	27
8	.95	4.43	6.00	7.18	8.12	9.03	9.78	10.5	11.1	11.7	12.2	12.7
	.99	7.50	9.9	11.7	13.2	14.5	15.8	16.9	17.9	18.9	19.8	21
9	.95	4.03	5.34	6.31	7.11	7.80	8.41	8.95	9.45	9.91	10.3	10.7
	.99	6.54	8.5	9.9	11.1	12.1	13.1	13.9	14.7	15.3	16.0	16.6
10	.95	3.72	4.85	5.67	6.34	6.92	7.42	7.87	8.28	8.66	9.01	9.34
	.99	5.85	7.4	8.6	9.6	10.4	11.1	11.8	12.4	12.9	13.4	13.9

(Continued)

TABLE B.4 (Continued)

Number of Variances each												
df for * each variance	1-α	2	3	4	5	6	7	8	9	10	11	12
12	.95	3.28	4.16	4.79	5.30	5.72	6.09	6.42	6.72	7.00	7.25	7.48
	.99	4.91	6.1	6.9	7.6	8.2	8.7	9.1	9.5	9.9	10.2	10.6
	.95	2.86	3.54	4.01	4.37	4.68	4.95	5.19	5.40	5.59	5.77	5.93
15	.99	4.07	4.9	5.5	6.0	6.4	6.7	7.1	7.3	7.5	7.8	8.0
	.95	2.46	2.95	3.29	3.54	3.76	3.94	4.10	4.24	4.37	4.49	4.59
	.99	3.32	3.8	4.3	4.6	4.9	5.1	5.3	5.5	5.6	5.8	5.9
30	.95	2.07	2.40	2.61	2.78	2.91	3.02	3.12	3.21	3.29	3.36	3.39
	.99	2.63	3.0	3.3	3.4	3.6	3.7	3.8	3.9	4.0	4.1	4.2
	.95	1.67	1.85	1.96	2.04	2.11	2.17	2.22	2.26	2.30	2.33	2.36
60	.99	1.96	2.2	2.3	2.4	2.4	2.5	2.5	2.6	2.6	2.7	2.7

Note: Reproduced with permission of the trustees of *Biometrika*.

*Equal group size (n) is assumed in the table; hence $df = n - 1$. If group sizes are not equal, then use the average group size (rounding off to the nearest integer) as the n .

TABLE B.5
Critical Values for Bryant-Paulson Procedure

Error df	Number of Covariates (C)	α	Number of Groups							
			2	3	4	5	6	7	8	10
3	1	.05	5.42	7.18	8.32	9.17	9.84	10.39	10.86	11.62
		.01	10.28	13.32	15.32	16.80	17.98	18.95	19.77	21.12
	2	.05	6.21	8.27	9.60	10.59	11.37	12.01	12.56	13.44
		.01	11.97	15.56	17.91	19.66	21.05	22.19	23.16	24.75
	3	.05	6.92	9.23	10.73	11.84	12.72	13.44	14.06	15.05
		.01	13.45	17.51	20.17	22.15	23.72	25.01	26.11	27.90
4	1	.05	4.51	5.84	6.69	7.32	7.82	8.23	8.58	9.15
		.01	7.68	9.64	10.93	11.89	12.65	13.28	13.82	14.70
	2	.05	5.04	6.54	7.51	8.23	8.80	9.26	9.66	10.31
		.01	8.69	10.95	12.43	13.54	14.41	15.14	15.76	16.77
	3	.05	5.51	7.18	8.25	9.05	9.67	10.19	10.63	11.35
		.01	9.59	12.11	13.77	15.00	15.98	16.79	17.47	18.60
5	1	.05	4.06	5.17	5.88	6.40	6.82	7.16	7.45	7.93
		.01	6.49	7.99	8.97	9.70	10.28	10.76	11.17	11.84
	2	.05	4.45	5.68	6.48	7.06	7.52	7.90	8.23	8.76
		.01	7.20	8.89	9.99	10.81	11.47	12.01	12.47	13.23
	3	.05	4.81	6.16	7.02	7.66	8.17	8.58	8.94	9.52
		.01	7.83	9.70	10.92	11.82	12.54	13.14	13.65	14.48
6	1	.05	3.79	4.78	5.40	5.86	6.23	6.53	6.78	7.20
		.01	5.83	7.08	7.88	8.48	8.96	9.36	9.70	10.25
	2	.05	4.10	5.18	5.87	6.37	6.77	7.10	7.38	7.84
		.01	6.36	7.75	8.64	9.31	9.85	10.29	10.66	11.28
	3	.05	4.38	5.55	6.30	6.84	7.28	7.64	7.94	8.44
		.01	6.85	8.36	9.34	10.07	10.65	11.13	11.54	12.22
7	1	.05	3.62	4.52	5.09	5.51	5.84	6.11	6.34	6.72
		.01	5.41	6.50	7.20	7.72	8.14	8.48	8.77	9.26
	2	.05	3.87	4.85	5.47	5.92	6.28	6.58	6.83	7.24
		.01	5.84	7.03	7.80	8.37	8.83	9.21	9.53	10.06
	3	.05	4.11	5.16	5.82	6.31	6.70	7.01	7.29	7.73
		.01	6.23	7.52	8.36	8.98	9.47	9.88	10.23	10.80
8	1	.05	3.49	4.34	4.87	5.26	5.57	5.82	6.03	6.39
		.01	5.12	6.11	6.74	7.20	7.58	7.88	8.15	8.58
	2	.05	3.70	4.61	5.19	5.61	5.94	6.21	6.44	6.82
		.01	5.48	6.54	7.23	7.74	8.14	8.48	8.76	9.23
	3	.05	3.91	4.88	5.49	5.93	6.29	6.58	6.83	7.23
		.01	5.81	6.95	7.69	8.23	8.67	9.03	9.33	9.84
10	1	.05	3.32	4.10	4.58	4.93	5.21	5.43	5.63	5.94
		.01	4.76	5.61	6.15	6.55	6.86	7.13	7.35	7.72
	2	.05	3.49	4.31	4.82	5.19	5.49	5.73	5.93	6.27
		.01	5.02	5.93	6.51	6.93	7.27	7.55	7.79	8.19
	3	.05	3.65	4.51	5.05	5.44	5.75	6.01	6.22	6.58
		.01	5.27	6.23	6.84	7.30	7.66	7.96	8.21	8.63

(Continued)

TABLE B.5 (Continued)

Error df	Number of Covariates (C)	α	Number of Groups							
			2	3	4	5	6	7	8	10
12	1	.05	3.22	3.95	4.40	4.73	4.98	5.19	5.37	5.67
		.01	4.54	5.31	5.79	6.15	6.43	6.67	6.87	7.20
	2	.05	3.35	4.12	4.59	4.93	5.20	5.43	5.62	5.92
		.01	4.74	5.56	6.07	6.45	6.75	7.00	7.21	7.56
	3	.05	3.48	4.28	4.78	5.14	5.42	5.65	5.85	6.17
		.01	4.94	5.80	6.34	6.74	7.05	7.31	7.54	7.90
14	1	.05	3.15	3.85	4.28	4.59	4.83	5.03	5.20	5.48
		.01	4.39	5.11	5.56	5.89	6.15	6.36	6.55	6.85
	2	.05	3.26	3.99	4.44	4.76	5.01	5.22	5.40	5.69
		.01	4.56	5.31	5.78	6.13	6.40	6.63	6.82	7.14
	3	.05	3.37	4.13	4.59	4.93	5.19	5.41	5.59	5.89
		.01	4.72	5.51	6.00	6.36	6.65	6.89	7.09	7.42
16	1	.05	3.10	3.77	4.19	4.49	4.72	4.91	5.07	5.34
		.01	4.28	4.96	5.39	5.70	5.95	6.15	6.32	6.60
	2	.05	3.19	3.90	4.32	4.63	4.88	5.07	5.24	5.52
		.01	4.42	5.14	5.58	5.90	6.16	6.37	6.55	6.85
	3	.05	3.29	4.01	4.46	4.78	5.03	5.23	5.41	5.69
		.01	4.56	5.30	5.76	6.10	6.37	6.59	6.77	7.08
18	1	.05	3.06	3.72	4.12	4.41	4.63	4.82	4.98	5.23
		.01	4.20	4.86	5.26	5.56	5.79	5.99	6.15	6.42
	2	.05	3.14	3.82	4.24	4.54	4.77	4.96	5.13	5.39
		.01	4.32	5.00	5.43	5.73	5.98	6.18	6.35	6.63
	3	.05	3.23	3.93	4.35	4.66	4.90	5.10	5.27	5.54
		.01	4.44	5.15	5.59	5.90	6.16	6.36	6.54	6.83
20	1	.05	3.03	3.67	4.07	4.35	4.57	4.75	4.90	5.15
		.01	4.14	4.77	5.17	5.45	5.68	5.86	6.02	6.27
	2	.05	3.10	3.77	4.17	4.46	4.69	4.88	5.03	5.29
		.01	4.25	4.90	5.31	5.60	5.84	6.03	6.19	6.46
	3	.05	3.18	3.86	4.28	4.57	4.81	5.00	5.16	5.42
		.01	4.35	5.03	5.45	5.75	5.99	6.19	6.36	6.63
24	1	.05	2.98	3.61	3.99	4.26	4.47	4.65	4.79	5.03
		.01	4.05	4.65	5.02	5.29	5.50	5.68	5.83	6.07
	2	.05	3.04	3.69	4.08	4.35	4.57	4.75	4.90	5.14
		.01	4.14	4.76	5.14	5.42	5.63	5.81	5.96	6.21
	3	.05	3.11	3.76	4.16	4.44	4.67	4.85	5.00	5.25
		.01	4.22	4.86	5.25	5.54	5.76	5.94	6.10	6.35
30	1	.05	2.94	3.55	3.91	4.18	4.38	4.54	4.69	4.91
		.01	3.96	4.54	4.89	5.14	5.34	5.50	5.64	5.87
	2	.05	2.99	3.61	3.98	4.25	4.46	4.62	4.77	5.00
		.01	4.03	4.62	4.98	5.24	5.44	5.61	5.75	5.98
	3	.05	3.04	3.67	4.05	4.32	4.53	4.70	4.85	5.08
		.01	4.10	4.70	5.06	5.33	5.54	5.71	5.85	6.08

(Continued)

TABLE B.5 (Continued)

Error df	Number of Covariates (C)	α	Number of Groups							
			2	3	4	5	6	7	8	10
40	1	.05	2.89	3.49	3.84	4.09	4.29	4.45	4.58	4.80
		.01	3.88	4.43	4.76	5.00	5.19	5.34	5.47	5.68
	2	.05	2.93	3.53	3.89	4.15	4.34	4.50	4.64	4.86
		.01	3.93	4.48	4.82	5.07	5.26	5.41	5.54	5.76
	3	.05	2.97	3.57	3.94	4.20	4.40	4.56	4.70	4.92
		.01	3.98	4.54	4.88	5.13	5.32	5.48	5.61	5.83
60	1	.05	2.85	3.43	3.77	4.01	4.20	4.35	4.48	4.69
		.01	3.79	4.32	4.64	4.86	5.04	5.18	5.30	5.50
	2	.05	2.88	3.46	3.80	4.05	4.24	4.39	4.52	4.73
		.01	3.83	4.36	4.68	4.90	5.08	5.22	5.35	5.54
	3	.05	2.90	3.49	3.83	4.08	4.27	4.43	4.56	4.77
		.01	3.86	4.39	4.72	4.95	5.12	5.27	5.39	5.59
120	1	.05	2.81	3.37	3.70	3.93	4.11	4.26	4.38	4.58
		.01	3.72	4.22	4.52	4.73	4.89	5.03	5.14	5.32
	2	.05	2.82	3.38	3.72	3.95	4.13	4.28	4.40	4.60
		.01	3.73	4.24	4.54	4.75	4.91	5.05	5.16	5.35
	3	.05	2.84	3.40	3.73	3.97	4.15	4.30	4.42	4.62
		.01	3.75	4.25	4.55	4.77	4.94	5.07	5.18	5.37

Source: Reproduced with permission of the trustees of *Biometrika*.

Appendix C

Power Tables

CONTENTS

Table C.1	Power of F Test at $\alpha = .05$, $u = 1$
Table C.2	Power of F Test at $\alpha = .05$, $u = 2$
Table C.3	Power of F Test at $\alpha = .05$, $u = 3$
Table C.4	Power of F Test at $\alpha = .05$, $u = 4$
Table C.5	Power of F Test at $\alpha = .10$, $u = 1$
Table C.6	Power of F Test at $\alpha = .10$, $u = 2$
Table C.7	Power of F Test at $\alpha = .10$, $u = 3$
Table C.8	Power of F Test at $\alpha = .10$, $u = 4$

NOTES

The quantity u refers to the degrees of freedom for the effect being tested. For a one way ANOVA with a levels we have $u = a - 1$. For a two way ANOVA with a levels for A and b levels for B, then $u = (a - 1)$ for the A main effect, $u = (b - 1)$ for the B main effect, and $u = (a - 1)(b - 1)$ for the interaction effect.

Group size is the assumed common number of subjects in each group. For two groups with unequal group sizes n_1 and n_2 , use the harmonic mean $2n_1n_2/(n_1 + n_2)$ to enter the table. For more than two groups, use the average group size to enter the table.

TABLE C.1
Power of *F* Test at $\alpha = .05, u = 1$

Group Size <i>n</i>	<i>f</i> (effect size)											
	.05	.10	.15	.20	.25	.30	.35	.40	.50	.60	.70	.80
4	05	06	06	07	09	11	13	16	23	30	39	48
5	05	06	07	08	11	13	16	20	29	39	50	61
6	05	06	07	09	12	15	20	24	35	47	60	71
7	05	06	08	10	14	18	23	28	41	55	68	79
8	05	06	08	11	15	20	26	32	47	62	75	85
9	05	07	09	12	17	22	29	36	52	68	80	89
10	05	07	09	13	18	25	32	40	57	73	85	93
11	05	07	10	14	20	27	35	44	62	77	88	95
12	05	07	10	15	22	29	38	47	65	81	91	97
13	05	07	11	16	23	32	41	51	70	84	93	98
14	05	08	11	17	25	34	44	54	73	87	95	98
15	06	08	12	18	26	36	47	57	76	89	96	99
16	06	08	12	19	28	38	49	60	79	91	97	99
17	06	08	13	20	30	40	52	63	82	93	98	*
18	06	08	14	21	31	42	54	66	84	94	98	
19	06	09	14	22	33	44	57	68	86	95	99	
20	06	09	15	23	34	46	59	70	88	96	99	
22	06	09	16	26	37	50	63	75	91	97		
24	06	10	17	28	40	54	67	78	93	98		
26	06	10	18	30	43	58	71	82	95	99		
28	06	11	19	32	46	61	74	84	96	99		
30	06	11	21	34	49	66	77	87	97			
32	06	12	22	36	51	67	80	89	98			
34	07	12	23	38	54	69	82	91	98			
36	07	13	24	40	56	72	84	92	99			
38	07	13	25	41	59	74	86	94	99			
40	07	14	27	43	61	77	88	95	99			
44	07	15	29	47	65	80	91	96				
48	07	16	31	50	69	84	93	97				
52	08	17	33	53	73	87	95	98				
56	08	18	36	57	75	89	96	99				
60	08	19	38	60	79	91	97	99				
64	08	20	40	62	81	93	98	*				
68	08	21	42	65	83	94	98					
72	09	22	44	68	85	95	99					
76	09	23	46	70	87	96	99					
80	09	24	48	72	89	97	99					
100	10	29	57	81	94	99						
140	13	39	72	92	99							
200	16	52	86	98								

TABLE C.3
Power of F Test at $\alpha = .05$, $u = 3$

[illegible]

TABLE C.5
Power of *F* Test at $\alpha = .10, u = 1$

Group Size <i>n</i>	<i>f</i> (effect size)											
	.05	.10	.15	.20	.25	.30	.35	.40	.50	.60	.70	.80
4	10	11	13	14	17	20	23	27	36	45	55	64
5	10	11	13	16	19	23	27	32	43	55	66	76
6	10	12	14	17	21	26	31	37	50	63	74	83
7	10	12	15	19	23	29	35	42	56	69	80	89
8	10	12	15	20	25	32	39	47	62	75	85	92
9	10	13	16	21	28	35	43	51	66	80	89	95
10	10	13	17	23	30	37	46	55	71	83	92	97
11	11	13	18	24	32	40	49	58	75	87	94	98
12	11	14	19	25	34	43	52	62	78	89	96	99
13	11	14	19	27	36	45	55	65	81	91	97	99
14	11	14	20	28	37	48	58	68	83	93	98	99
15	11	15	21	29	39	50	60	70	86	95	98	*
16	11	15	22	31	41	52	63	73	88	96	99	
17	11	15	23	32	43	54	65	75	89	97	99	
18	11	16	23	33	45	56	68	77	91	97	99	
19	11	16	24	34	46	58	70	79	92	98	*	
20	11	16	25	36	48	60	72	81	93	98		
22	11	17	26	38	51	64	75	84	95	99		
24	12	18	28	40	54	67	78	87	96	99		
26	12	19	29	43	57	70	81	89	97	*		
28	12	19	31	45	60	73	84	91	98			
30	12	20	32	47	62	76	86	93	99			
32	12	21	34	49	65	78	89	94	99			
34	12	21	35	51	67	80	90	95	99			
36	13	22	36	53	69	82	91	96	*			
38	13	23	38	55	71	84	92	97				
40	13	24	39	57	73	85	93	97	*			
44	13	25	42	60	77	88	95	98				
48	14	26	44	63	80	91	96	99				
52	14	28	47	66	82	92	97	99				
56	14	29	49	69	85	94	98	*				
60	15	30	51	72	87	95	99					
64	15	31	53	74	89	96	99					
68	16	33	56	76	90	97	99					
72	16	34	58	78	92	98	99					
76	16	35	59	80	93	98	*					
80	17	36	61	82	94	99						
100	18	42	70	89	97							
140	22	53	82	96	99							
200	27	65	92	99								

TABLE C.7
Power of F Test at $\alpha = .10$, $u = 3$

[illegible]

References

- Agresti, A. (1990). *Statistical methods for social science*. Englewood Cliffs, NJ: Prentice Hall.
- Anderson, N. H. (1963). Comparison of different populations: Resistance to extinction and transfer. *Psychological Bulletin*, 70, 162–179.
- Anscombe, F. (1973). Graphs in statistical analysis. *American Statistician*, 27, 11–21.
- Barcikowski, R. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, 6, 267–285.
- Barcikowski, R., & Robey, R. (1984). Decisions in a single group repeated measures analysis: Statistical tests and three computer packages. *American Statistician*, 38, 248–250.
- Becker, B. (1987). Applying tests of combined significance in meta-analysis. *Psychological Bulletin*, 102, 164–171.
- Bloom, B. (1984). The 2-sigma problem: The search for methods of group instruction as effective as one-to-one. *Educational Researcher*, 13, 4–16.
- Bock, R. D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Box, G. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems: II. Effect of inequality of variance and of correlations between errors in the two way classification. *Annals of Mathematical Statistics*, 25, 484–498.
- Brown, M. B., & Forsythe, A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, 69, 364–367.
- Bryan, T. (1974). Peer popularity of learning disabled children. *Journal of Learning Disabilities*, 7, 31–35.
- Bryant, J. L., & Paulson, A. S. (1976). An extension of Tukey's method of multiple comparisons to experimental designs with random concomitant variables. *Biometrika*, 63, 631–638.
- Bryk, A. S. (1977). Evaluating program impact: A time to cast away stones, a time to gather stones together. *New Directions for Program Evaluation*, 1, 31–58.
- Bryk, A. S. (1992) Hierarchical linear models: applications and data analysis methods (1st edition). Thousand Oaks, CA: Sage Publications, Inc.
- Bryk, A. S., & Weisberg, H. I. (1977). Use of the nonequivalent control group design when subjects are growing. *Psychological Bulletin*, 85, 950–962.
- Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education*, 8, 158–233.
- Carlson, J., & Timm, N. (1974). Analysis of non-orthogonal fixed effects designs. *Psychological Bulletin*, 81, 563–570.
- Chase, C. (1986). Essay test scoring: Interaction of relevant variables. *Journal of Educational Measurement*, 23, 33–41.

- Cobb, G. (1987). Introductory textbooks: A framework for evaluation. *Journal of the American Statistical Association*, 82, 321–339.
- Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, 13, 261–281.
- Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, 70, 426–443.
- Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. (1973). Eta squared and partial eta squared in fixed factor designs. *Educational and Psychological Measurement*, 33, 107–112.
- Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 15, 1304–1312.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Collier, R., Baker, F., Mandeville, C. K., & Hayes, T. (1967). Estimates of test size for several test procedures on conventional variance ratios in the repeated measures design. *Psychometrika*, 32, 339–353.
- Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, 19, 15–18.
- Cook, R. D., & Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Cradler, J., & Goodwin, D. (1971). Conditioning of verbal behavior as a function of age, social class and type of reinforcement. *Journal of Educational Psychology*, 62, 279–284.
- Cronbach, L. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, 30, 116–127.
- Cronbach, L., & Snow, R. (1969). *Individual differences in learning ability as a function of instructional variables*. Unpublished report, School of Education, Stanford University, Stanford, CA.
- Cronbach, L., & Snow, R. (1977). *Aptitudes and instructional methods*. New York: Irvington Press.
- Crowder, R. (1975). An investigation of the relationship between social IQ and vocational evaluation ratings with an adult trainable mental retardate work activity center population. Unpublished doctoral dissertation, University of Cincinnati, OH.
- Crystal, G. (1988). The wacky, wacky world of CEO pay. *Fortune*, 117, 68–78.
- Dance, K., & Neufeld, R. (1988). Aptitude treatment interaction research in the clinical setting: A review of attempts to dispel the “Patient Uniformity” myth. *Psychological Bulletin*, 104, 192–213.
- Daniels, R., & Stevens, J. (1976). The interaction between the internal-external locus of control and two methods of college instruction. *American Educational Research Journal*, 13, 103–113.
- Davidson, M. L. (1972). Univariate versus multivariate tests in repeated measures experiments. *Psychological Bulletin*, 77, 446.

- Dizney, H., & Gromen, L. (1967). Predictive validity and differential achievement on three MLA Comparative Foreign Language tests. *Educational and Psychological Measurement*, 27, 1127–1130.
- Draper, N., & Smith, H. (1981). *Applied regression analysis*. New York: Wiley.
- Dunnett, C. W. (1980). Pairwise multiple comparisons in the homogeneous variance, unequal sample size cases. *Journal of the American Statistical Association*, 75, 789–795.
- Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatments with a control. *Journal of the American Statistical Association*, 50, 1096–1121.
- Elashoff, J. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal*, 6, 383–401.
- Elashoff, J. (1981). Data for the panel session in software for repeated measures analysis of variance. *Proceedings of the Statistical Computing Section*, American Statistical Association.
- Feshbach, S., Adelman, H., & Williamson, F. (1977). Prediction of reading and related academic problems. *Journal of Educational Psychology*, 69, 299–308.
- Frane, J. (1976). Some simple procedures for handling missing data in multivariate analysis. *Psychometrika*, 41, 409–415.
- Games, P., & Howell, J. K. (1976). Pairwise multiple comparison procedures with unequal *ns* and/or variances: A Monte Carlo study. *Journal of Educational Statistics*, 1, 113–125.
- Glass, G., & Hopkins, K. (1984). *Statistical methods in education and psychology*. Englewood Cliffs, NJ: Prentice Hall.
- Glass, G., Peckham, P., & Sanders, J. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, 42, 237–288.
- Glasnapp, D., & Poggio, J. (1985). *Essentials of statistical analysis for the behavioral sciences*. Columbus, OH: Charles Merrill.
- Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN*. Multilevel Models Project, University of London.
- Greenhouse, S., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, 24, 95–112.
- Hand, D. J., & Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures*. London: Chapman and Hall.
- Harrington, S. (1968). *Sequencing organizers in meaningful verbal learning* (Research Paper No. 10). Boulder: University of Colorado, Laboratory of Educational Research.
- Hays, W. (1963). *Statistics for psychologists*. New York: Holt, Rinehart and Winston.
- Hays, W. (1981). *Statistics* (3rd ed.). New York: Holt, Rinehart and Winston.
- Hayter, A. (1984). A proof of the conjecture that the Tukey-Kramer multiple comparison procedure is conservative. *Annals of Statistics*, 12, 61–75.
- Herzberg, P. A. (1969). The parameters of cross validation. *Psychometrika* (Monogr. Suppl., No. 16).
- Hoaglin, D., & Welsch, R. (1978). The hat matrix in regression and ANOVA. *American Statistician*, 32, 17–22.

- Holland, B. S., & Copenhaver, M. D. (1988). Improved Bonferroni type multiple testing procedures. *Psychological Bulletin*, 104, 145–149.
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6, 65–70.
- Hox, J. J. (2002). *Multilevel analysis: techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Huberty, C. J. (1987). On statistical testing. *Educational Researcher*, 16, 4–9.
- Huberty, C. J. (1989). Problems with stepwise methods—Better alternatives. In B. Thompson (Ed.), *Advances in social science methodology* (Vol. I, pp. 43–70), Greenwich, CT: JAI Press.
- Huck, S., & Bounds, W. (1972). Essay grades: An interaction between graders handwriting clarity and the neatness of examination papers. *American Educational Research Journal*, 9, 279–283.
- Huck, S., & McLean, R. A. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin*, 82, 511–518.
- Huck, S., Cormier, W., & Bounds, W. (1974). *Reading statistics and research*. New York: Harper and Row.
- Huitema, B. (1980). *The analysis of covariance and alternatives*. New York: Wiley.
- Huynh, H., & Feldt, L. (1970). Conditions under which mean square ratios in repeated measures designs have exact distributions. *Journal of the American Statistical Association*, 65, 1582–1589.
- Huynh, H., & Feldt, L. (1976). Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split plot designs. *Journal of Educational Statistics*, 1, 69–82.
- Jennings, E. (1988). Models for pretest-posttest data: Repeated measures ANOVA revisited. *Journal of Educational Statistics*, 13, 273–280.
- Johnson, P. O., & Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs*, 1, 57–93.
- Johnson, R., & Wichern, D. (2002). *Applied multivariate statistical analysis* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.
- Jones, L. V., Lindzey, G., & Coggeshall, P. (Eds.) (1982). *An Assessment of Research-Doctorate Programs in the United States: Social and Behavioral Sciences*, (Washington, DC: National Academy Press).
- Kenny, D., & Judd, C. (1986). Consequences of violating the independence assumption in analysis of variance. *Psychological Bulletin*, 99, 422–431.
- Keppel, G. (1983). *Design and analysis: A researcher's handbook*. Englewood Cliffs, NJ: Prentice Hall.
- Kerlinger, F., & Pedhazur, E. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart & Winston.
- Keselman, H. J., & Keselman, J. C. (1988). Repeated measures multiple comparison procedures: Effects of violating multisample sphericity in unbalanced designs. *Journal of Educational Statistics*, 13, 215–226.

- Keselman, H. J., Murray, R. M., & Rogan, J. (1976). *Effect of very unequal group sizes on Tukey's multiple comparison test*. Paper presented at the annual meeting of the American Educational Research Association, 1975, Washington, DC.
- Keselman, H. J., Rogan, J., Mendoza, J., & Breen, L. (1980). Testing the validity conditions of repeated measures F tests. *Psychological Bulletin*, 87, 479–481.
- Kirk, R. (1982). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Brooks-Cole.
- Kreft, I., & de Leeuw, J. (1998). *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage Publications, Inc.
- Levin, J., McCormick, C., Miller, G., Berry, J., & Presley, M. (1982). Mnemonic versus nonmnemonic vocabulary learning strategies for children. *American Educational Research Journal*, 19, 121–136.
- Light, R., & Pillimer, D. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS Institute, Inc.
- Longford, N. T. (1988). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 297–310). Orlando, FL: Academic Press.
- Lord, F. M. (1969). Statistical adjustments when comparing pre-existing groups. *Psychological Bulletin*, 70, 162–179.
- Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mallows, C. L. (1973). Some comments on C_p . *Technometrics*, 15, 661–676.
- Marwit, S., & Neumann, G. (1974). Black and white children's comprehension of standard and nonstandard English passages. *Journal of Educational Psychology*, 66, 324–332.
- Maxwell, S. (1980). Pairwise multiple comparison procedures in repeated measures designs. *Journal of Educational Statistics*, 5, 269–287.
- Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd edition). Mahwah, NJ: Lawrence Erlbaum Associates.
- Moore, D., & McCabe, G. (1989). *Introduction to the practice of statistics*. New York: Freeman.
- Morrison, D. F. (1976). *Multivariate statistical methods*. New York: McGraw-Hill.
- Morrison, D. F. (1983). *Applied linear statistical methods*. Englewood Cliffs, NJ: Prentice Hall.
- Myers, J. (1979). *Fundamentals of experimental design*. Boston: Allyn and Bacon.
- Myers, J., & Well, A. (1991). *Research design and statistical analysis*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Myers, R. (1990). *Classical and modern regression with applications* (2nd ed.). Boston, MA: Duxbury Press.
- Novince, L. (1977). *The contribution of cognitive restructuring to the effectiveness of behavior rehearsal in modifying social inhibition in females*. Unpublished doctoral dissertation, University of Cincinnati.

- Nunnally, J. (1978). *Psychometric theory*. New York: McGraw-Hill.
- O'Brien, R., & Kaiser, M. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin*, 97, 316–333.
- O'Grady, K. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin*, 92, 766–777.
- Overall, J., & Spiegel, D. (1969). Concerning the least squares analysis of experimental data. *Psychological Bulletin*, 72, 311–322.
- Park, C., & Dudycha, A. (1974). A cross validation approach to sample size determination for regression models. *Journal of the American Statistical Association*, 69, 214–218.
- Pedhazur, E. (1982). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston.
- Porter, A. (1967). *The effects of using fallible variables in the analysis of covariance*. Unpublished doctoral dissertation, University of Wisconsin, Madison.
- Pukulski, J. (1970). Effects of reinforcement on word recognition. *The Reading Teacher*, 23, 516–522.
- Raudenbush, S. W. (1984). *Applications of a hierarchical linear model in educational research*. Unpublished doctoral dissertation, Harvard University.
- Raudenbush, S. W., & Byrk, A. S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, 12, 241–269.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd edition). Thousand Oaks, CA: Sage Publications, Inc.
- Raudenbush, S. W., Bryk, A., Cheong, Y. F., & Congdon, R. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.
- Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. Cook & D. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–205). Chicago: Rand McNally.
- Rogosa, D. (1977). *Some results for the Johnson-Neyman technique*. Doctoral dissertation, Stanford University, Stanford, CA.
- Rogosa, D. (1980). Comparing non-parallel regression lines. *Psychological Bulletin*.
- Rosenthal, R., & Rosnow, R. (1984). *Essentials of behavioral research*. New York: McGraw-Hill.
- Rounet, H., & Lepine, D. (1970). Comparison between treatments in a repeated measures design: ANOVA and multivariate methods. *British Journal of Mathematical and Statistical Psychology*, 213, 147–163.
- Sarachan-Diely, A. (1985). Written narrative of deaf and hearing students: Story recall and inference. *Journal of Speech and Hearing Research*, 28, 151–159.
- SAS Institute Inc. (1999). *SAS/STAT User's Guide*, Version 8, 3 volume set. Cary, NC.
- Scariano, S., & Davenport, J. (1987). The effects of violations of independence assumption in the one-way ANOVA. *American Statistician*, 41, 123–129.
- Schutz, W. (1977). *Leaders of schools: FIRO theory applied to administrators*. LaJolla, CA: University Associates.
- Shiffler, R. (1988). Maximum z scores and outliers. *American Statistician*, 42, 79–80.

- Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of Educational and Behavioral Statistics*, 23, 323–355.
- Singer, J. D., & Willett, J. B. (1988). *Opening up the black box of recipe statistics: Putting the data back into data analysis*. Paper presented at the annual meeting of the American Educational Research Association, April, New Orleans, LA.
- Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.
- Smith, S., Jones, L. & Waugh, M. (1986). Production and evaluation of interactive videodisc lessons in laboratory instruction. *Journal of Computer-Based Instruction*, 13, 117–121.
- Snijders, T. & Bosker, R. (1999). *Multilevel Analysis*. Thousand Oaks, CA: Sage Publications, Inc.
- SPSS Inc. (2003). *SPSS Base 12.0 User's Guide*. Chicago.
- Stein, C. (1960). Multiple regression. In I. Olkin (Ed.), *Contributions to probability and statistics, essays in honor of Harold Hotelling*. Stanford, CA: Stanford University Press.
- Stevens, J. P. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Lawrence Erlbaum Associates.
- Stoloff, P. H. (1967). *An empirical evaluation of the effects of violating the assumption of homogeneity of covariance for the repeated measures design of the analysis of variance* (Tech. Rep.). College Park: University of Maryland.
- Thorndike, R. L., & Hagen, E. (1977). *Measurement and evaluation in psychology and education*. New York: Wiley.
- Tomarkin, J. J., & Serlin, R. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, 99, 90–99.
- Tuckman, B., Steber, J., & Hyman, R. (1979). Judging the effectiveness of teaching styles: The perceptions of principals. *Educational Administration Quarterly*, 15, 104–115.
- Weisberg, S. (1985). *Applied linear regression*. New York: Wiley.
- Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika*, 38, 330–336.
- Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin*, 86, 168–174. Arnold, C. L. (1992). An introduction to hierarchical linear models. *Measurement and Evaluation in Counseling and Development*, 25, 58–90.

Answers to Selected Exercises

CHAPTER 1

1. (a) The \$22,000 figure was misleading because of the extreme salaries of \$70,000 and \$250,000. Recall that the mean is very sensitive to extreme values.
 (b) The median should have been used. It is essentially unaffected by extreme values. The median for this set of data is \$15,000, and indicates where most of the salaries are concentrated.
3. She should not be concerned, since considerable research has shown that a violation of the normality assumption has little effect on the Type I error rate.
5. (a) $\sum c x_i = \sum 3x_i = \sum x_i = 3(5 + 8 + 1 + 7) = 63$
 (b) First, note that the mean for the scores cx_1, cx_2, \dots, cx_n is given by

$$\bar{x}_{cx} = \frac{cx_1 + cx_2 + \dots + cx_n}{n} = \frac{c(x_1 + x_2 + \dots + x_n)}{n} = c\bar{x}$$

where \bar{x} is the mean for x_1, x_2, \dots, x_n .

Now, using the definitional formula for variance, we have

$$\begin{aligned} s_{cx}^2 &= \frac{\sum (cx_i - c\bar{x})^2}{n-1} = \frac{\sum [c(x_i - \bar{x})]^2}{n-1} = \frac{\sum c^2(x_i - \bar{x})^2}{n-1} = \frac{c^2 \sum (x_i - \bar{x})^2}{n-1} \\ &= c^2 s_x^2, \text{ as was to be proved.} \end{aligned}$$

- (c) The grand mean for the groups is 6.3. Therefore,

$$\begin{aligned} \sum 10(\bar{x}_i - \bar{x})^2 &= \sum 10(\bar{x}_i - 6.3)^2 \\ &= 10[(4.1 - 6.3)^2 + (8.5 - 6.3)^2] = 96.8 \end{aligned}$$

7. (a) The correlation for all 14 data points is .587, indicating a moderate relationship between height and weight.
 (b) The outlier is subject 10, whose weight of only 115 lbs is very unusual for someone over 6 ft tall.
 (c) The correlation without subject 10 is now .867, indicating that there is indeed a strong relationship between height and weight.

9. (a) The 95% confidence interval for the first study is given by $4 \pm 2.101(.57)$, or $(2.8, 5.2)$, while the 95% confidence interval for the second study is given by $4 \pm 1.96(1.74)$, or $(.59, 7.41)$. The null hypothesis of equal population means is rejected in both cases, since 0 is *not* in either interval.
- (b) We can be confident of clinical significance in the first study since the interval is indicating that the difference in the population means is at least 2.8, that is, greater than 2. We cannot be confident of clinical significance in the second study since the interval indicates the population mean difference could be as small as .59.

CHAPTER 2

1. Overall $\alpha = 1 - (1 - .05)^{21} = 1 - .34 = .66$
3. (a) $df_b = 2$, $df_w = 27$. Therefore, the critical value at the .05 level is 3.35.
 (b) $df_b = 3$, $df_w = 76$. Therefore, the critical value at the .10 level is 2.17.
 (c) $df_b = 4$, $df_w = 35$. Therefore, the critical value at the .01 level is 3.9.
5. Using a calculator, we obtain first the means and variances for each group:

	GROUP 1	GROUP 2	GROUP 3	GROUP 4
\bar{x}_i	4	9	4.8	6.5
s_i^2	3.33	4	3.7	3

Now, sum of squares within is given by

$$\begin{aligned} SS_w &= (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2 \\ &= 3(3.33) + 2(4) + 4(3.7) + 3(3) = 41.79 \end{aligned}$$

$$MS_w = SS_w / (N - k) = 41.79 / 12 = 3.48$$

$$\begin{aligned} SS_b &= 4(4 - 5.81)^2 + 3(9 - 5.81)^2 + 5(4.8 - 5.81)^2 + 4(6.5 - 5.81)^2 \\ &= 50.64 \end{aligned}$$

$$MS_b = SS_b / (k - 1) = 50.64 / 3 = 16.88$$

Therefore, $F = MS_b / MS_w = 16.88 / 3.48 = 4.85$

The critical value at the .05 level is 3.49. Thus, we reject and conclude there is an overall difference among the groups.

7. Recall that the formula for obtaining the intervals is

$$(\bar{x}_i - \bar{x}_j) \pm q_{\alpha; k; N-k} \sqrt{MS_w / n},$$

where q is studentized range statistic (Table D), MS_w is the error term from the ANOVA, and n is the common number of subjects per group.

$$q\sqrt{MS_w/n} = 3.312\sqrt{22.35/15} = 4.043$$

Now, we obtain the confidence intervals:

GROUPS	CRITICAL VALUE	CONFIDENCE INTERVALS
$\bar{x}_1 - \bar{x}_2 = -1.7$	4.043	(-5.743, 2.343)
$\bar{x}_1 - \bar{x}_3 = 1.5$	4.043	(-2.543, 5.543)
$\bar{x}_1 - \bar{x}_4 = -3.1$	4.043	(-7.143, .943)
$\bar{x}_2 - \bar{x}_3 = 3.2$	4.043	(-.843, 7.243)
$\bar{x}_2 - \bar{x}_4 = -1.4$	4.043	(-5.433, 2.643)
$\bar{x}_3 - \bar{x}_4 = -4.6$	4.043	(-8.643, -.557)

Since the intervals for the first 5 paired comparisons all cover 0, none of these are significant. Only the last paired comparison is significant, since that interval does not cover 0, that is, 0 is not a likely value for $\mu_3 - \mu_4$.

9. The estimate of the contrast is

$$\begin{aligned}
 L_2 &= (5.6 + 7.3)/2 - (8.1 + 4.2)/2 = .30 \\
 \sum c_i^2 / n_i &= (.5)^2 / 10 + (.5)^2 / 8 + (-.5)^2 / 11 + (-.5)^2 / 13 \\
 &= .098 \\
 F &= \frac{(.30)^2 / .098}{8.7} < 1
 \end{aligned}$$

Since we have more error variation than effect variation, the contrast is clearly not significant.

11. O'Grady's statement relates to the restriction of range phenomenon you encountered when studying the Pearson correlation in your introductory statistics course. In this case there would undoubtedly be the least amount of variance in heart efficiency to account for in a population of runners (more homogeneous), while a random sample of the American adult population is much more heterogeneous and therefore the potential of accounting for more variance.
13. The form of the control lines for running the analysis is identical to that presented in the chapter. To obtain both the Scheffe and Tukey intervals in one run simply insert in the MEANS statement:

MEANS REGION/SCHEFFE TUKEY;

where REGION is the name I have given to the grouping variable.

(a) There is a significant overall difference at the .05 level, since from the printout we have $F = 3.38$, $p = .0285$.

(b) There are no significant pairwise differences found with the Scheffé procedure, while a significant pairwise difference, between Groups 1 and 4, is found with the Tukey procedure.

(c) The Scheffé is a more conservative procedure than Tukey, and is not as powerful for detecting pairwise differences.

15. (a) The 5 comparisons are given schematically:

	KEYWORD	EXPERIENTIAL	PICTURE	CONTROL
$L1$	1	0	0	-1
$L2$	1	0	-1	0
$L3$	1	-1	0	0
$L4$	0	0	1	-1
$L5$	0	1	0	-1

(b) First of all, note that the set of 5 comparisons must have dependencies since there are only 3 degrees of freedom between, and hence at most 3 independent comparisons. If we compute the sum of products for $L1$ and $L2$ we find

$$1(1) + 0(0) + 0(-1) + (-1)(0) = 1$$

Therefore, there is a dependency for $L1$ and $L2$.

(c) Since the group sizes are equal ($n = 16$), the error term (MS_w) for each contrast is simply the average of the group variances. Therefore,

$$MS_w = \frac{(22.9)^2 + (27)^2 + (23.1)^2 + (25.6)^2}{4} = 610.6$$

Recall from page 68 that if the group sizes are equal, then the F statistic for testing a contrast for significance is

$$F = \frac{n\hat{L}^2 / \sum c_i^2}{MS_w}$$

Keyword Versus Control

$$F = \frac{16(72.3 - 48.7)^2 / 2}{610.6} = 7.2972 \Rightarrow t = 2.701$$

Keyword Versus Picture

$$F = \frac{16(72.3 - 42.4)^2 / 2}{610.6} = 11.713 \Rightarrow t = 3.42$$

Keyword Versus Experiental

$$F = \frac{16(72.3 - 36.2)^2 / 2}{610.6} = 17.074 \Rightarrow T = 4.132$$

17. (a) The null hypothesis is $\mu_1 = \mu_2 = \mu_3$. It is rejected at the .10 level since $F = 3.115$ and $p = .053$.
 (b) The Levene test is not significant at the .05 level, since $p = .766$.
 (c) For the Tukey procedure at the .10 level, only Groups 1 and 3 are significantly different.

SELECTED PRINTOUT FROM SPSS FOR WINDOWS

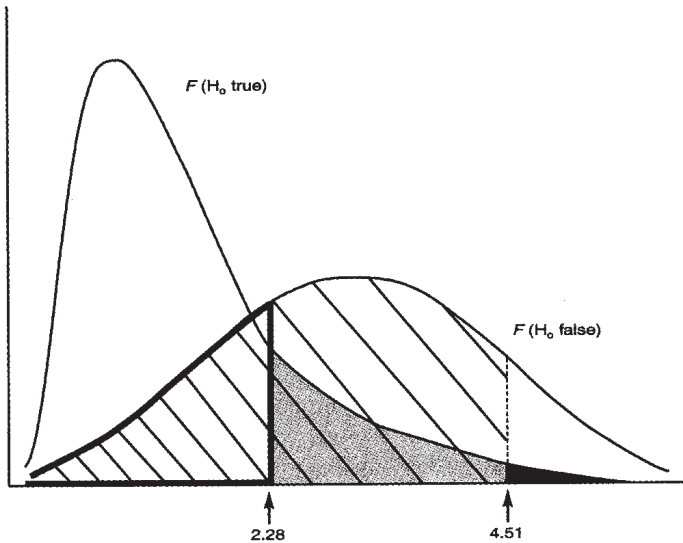
Insert art from p. 404 of previous edition. (x 41.5 pi)

19. (a) $1 - (1 - \alpha')^3 = 1 - (1 - 3\alpha' + 3\alpha'^2 - \alpha'^3) = 3\alpha' - 3\alpha'^2 + \alpha'^3$
 (b) $3\alpha' = 3(.01) = .03$
 $3\alpha' - 3\alpha'^2 + \alpha'^3 = 3(.01) - 3(.01)^2 + (.01)^3 = .0297$

What we have shown is that the two quantities are approximately the same for small α .

CHAPTER 3

1.



We have shown that as Type I error decreases (from light shaded area to dark shaded), Type II error increases (from boldfaced lined area to the total lined area).

3. In doing a two-tail test, say at .05, the alpha level is divided into two equal portions of .025. Thus, in effect we are working at a more severe alpha level, and therefore we will have less power for the two tailed test.
5. Using the formula for effect size, $\hat{d} = t\sqrt{(1/n_1 + 1/n_2)}$, we obtain

(a)

Study	\hat{d}_i
1	.69
2	.72
3	.56
4	.37
5	.89
6	.67
8	.48
9	.93

(b) The vast majority of the studies (8 of 10) show medium to large effect sizes, which would undoubtedly be of practical significance. Yet in 7 of the 8 cases, significance was not found because of a power problem (small to very small group sizes). There is systematic evidence to document the superiority of the combined treatment.

7. To estimate power we first find the estimated effect size \hat{f} , from $\hat{f} = \sqrt{(k-1)F / N} = \sqrt{4(2.03) / 125} = .255$
Now, using Table C.4, with $f = .25$ and $n = 25$, we find that power = .58.
To obtain power at $\alpha = .10$, we use Table C.8 and find that power = .71.
9. (A) $\hat{f} = \sqrt{3(5.61) / 800} = .145$. Now, using Table C.3 with $f = .15$ and an average group size of 200 we find that power = .96.
(b) These results do not appear to have any practical significance. First, the effect size is small. Secondly, look at the size of the mean differences (for a scale which has a range from 10 to 50). The mean differences for all pairs of groups, except Jewish and Protestant 2, are about 2 or less. These are trivial differences on a scale with a range of 40.
11. Below is the PASS 6.0 printout.
Power does not become adequate for any of the sample sizes. Note that at the .05 level the power is only .34 even with 120 subjects per group! Also, even at the .10 level, the power is just .4626 with 120 subjects per group.

SELECTED PRINTOUT FROM PASS 6.0

Insert art from p. 406 of previous edition

CHAPTER 4

1. A fourth advantage of a factorial design is that it can help to increase the generalizability of results. For example, suppose we had compared 3 treatments in a one way ANOVA and found a significant difference. Someone then says to us that the relative efficacy of treatments might depend on the sex of the subjects, and we run a factorial ANOVA (sex by treatments) to check this out. If we had adequate power and the interaction effect is not significant, we can generalize our results.
3. The control lines for running Problem 2 are as follows:

```
DATA TWOWAY;
INPUT AGE TREAT DEP @@;
CARDS;
1 1 21 1 1 27 1 1 23 1 1 28 1 1 20
1 2 24 1 2 32 1 2 30 1 2 35 1 2 32
1 3 19 1 3 30 1 3 27 1 3 20 1 3 21
2 1 18 2 1 25 2 1 27 2 1 20 2 1 23
2 2 24 2 2 16 2 2 18 2 2 19 2 2 20
2 3 34 2 3 28 2 3 21 2 3 30 2 3 29
PROC PRINT;
PROC GLM;
CLASS AGE TREAT;
MODEL DEP = AGE TREAT AGE *TREAT;
MEANS AGE TREAT AGE*TREAT;
```

Selected printout from the above run is given below:

SOURCE	DF	TYPE I SS	F VALUE	PR > F	TYPE III SS
AGE	1	45.63333	2.82	.1063	45.63333
TREAT	2	37.80000	1.17	.3284	37.80000
AGE*TREAT	2	334.06667	10.31	.0006	334.06667

Notice that the Type I and Type III sums of squares are the same, as they always will be for equal cell size factorial ANOVA. The F values will also be the same, and are not repeated twice here. If we had decided to test each effect at the .05 level, then only the interaction effect is significant, since only that p value is less than .05.

5. (a) $f_B = \sqrt{2 \cdot 5.57 / 24} = .681$

The n needed to enter the table is

$$n_B = [(N - rc) / c] + 1 = [(24 - 6) / 3] + 1 = 7$$

Now, using Table C.6 and $f = .70$ (since our estimated effect size is very close to this value), we find that power is .86. Actually, if we had interpolated power would be slightly less, but the main point here is that power for detecting the reinforcement main effect was quite good.

(b) $\hat{f}_{AB} = \sqrt{2 \cdot 1.87 / 24} = .395$

The n needed to enter the table is

$$n_{AB} = \frac{(N - rc)}{(r - 1)(c - 1) + 1} + 1 = \frac{(24 - 6)}{(2 - 1)(3 - 1) + 1} + 1 = 7$$

(c) Given that Pukulski had less than a 50% chance of detecting a large interaction effect, the study should be replicated with larger sample size for more adequate power. Using Table C.6, we see that for $f = .40$, power is only .45.

7. (a) For each dependent variable in a two-way design, there are 3 statistical tests (2 main effects and interaction effects). Since 5 two-way ANOVAs were done, this means $5(3) = 15$ statistical tests were done.

(b) Upper bound on overall α is $1 - (1 - .05)^{15} = 1 - .463 = .537$.

(c) The investigator should be quite cautious, since the probability of at least a few spurious rejections is very high.

9. (a) We present below the rearranged means along with the row, column, and grand means, and the estimated interaction effects:

	TREATMENT		ROW MEANS	
	6.5 (-.1667)	7.8333 (.5)	8.3333 (-.3333)	7.5556
AGE	8.8333 (.1667)	8.8333 (-.5)	11 (.3333)	9.5556
COLUMN MEANS	7.6667	8.3333	9.6667	8.5556 (GRAND MEAN)

$$SS = 6[2(.0278) + 2(.25) + 2(.1111)] = 6(.7778) = 4.6668$$

- (b) We follow the same process as above for calculating the sex by age sum of squares:

	AGE		ROW MEANS	
	7.5556 (-.0556)	9.6667 (.0556)	8.6112	
SEX	7.5556 (.0556)	9.4444 (-.0556)	8.5000	
COLUMN MEANS	7.5556	9.5556	8.5556	(GRAND MEAN)

$$SS = 9[4(.00309)] = .1112$$

(c) $SS_{age} = 18 [(7.5556 - 8.5556)^2 + (9.5556 - 8.5556)^2] = 36$

11. The n to enter Cohen's tables for the treatment main effect is given by $n = [(90-9)/3] + 1 = 28$. The power at .05 is .52 and power at .10 is .65 (here $u = 2$, since there are 2 df for treatment). Thus, power is still not quite adequate, even at the .10 level of significance. For the interaction effect, the n to enter the table is $n = [(90-9)/(4+1)] + 1 = 17$ (approx.). The degrees of freedom for interaction here is $(3-1)(3-1) = 4$, which recall is u in Cohen's tables. Thus, power = .40 at .05 and .54 at .10. Assuming power would increase roughly by the same amount (.14), in going from $\alpha = .10$ to $\alpha = .15$, we estimate that power would be about .68 at $\alpha = .15$.

13. (a) The significant dynamism and warmth and acceptance interaction effects indicate that the principals rate more versus less effective teachers differentially on each of this traits, depending on the level of schooling.
(b) The cell means for dynamism and warmth and acceptance are

	DYNAMISM		WARMTH & ACCEPT.	
	MORE EFF	LESS EFF	MORE EFF	LESS EFF
ELEM	25.7	28.9	39.3	23.9
INTERM	27.9	22.8	35.6	26.5
SENIOR	28.2	17.6	31.7	26.7

Note that the means for dynamism increased (for the more effective teachers) as the school level of the principal increases, and decreased for the less effective teachers, as the authors hypothesized. Recall that an interaction can be thought of as a difference in the differences, and here those differences are -3.2 , 5.1 , and 10.6 .

Regarding warmth and acceptance, the difference in means for more and less effective teachers is sharpest for the elementary principals and decreases in size as the level of the principal increases (as the authors had hypothesized). The differences are 15.4 , 9.1 , and 5 .

(c) Basically half of their hypotheses were confirmed, that is, that warmth and acceptance and dynamism would be important in distinguishing more versus less effective teachers. However, they also hypothesized that creativity would discriminate for elementary school principals, while organized demeanor would be important for intermediate and high school principals, and neither of these was confirmed. As a matter of fact, the discrepancy between more versus less effective teachers on creativity is sharpest for the senior-level principals. On organized demeanor the means for elementary, intermediate, and high school principals are 34.8 , 36.8 , and 36.3 .

(d) To check which pairs of means on organized demeanor are significantly different one should use the Tukey procedure. You will find that there are no significant differences.

15. (a) From the printout the following effects are significant at the $.01$ level: SEX ($F = 22.067$, $p = .000$), TREAT ($F = 8.685$, $p = .001$) and SEX * TREAT ($F = 6.034$, $p = .005$)
(b) From the marginal means, it is clear that males (if males are coded as 1) did better than females, and that Treatment 2 was the best (assuming higher is better). However, the interaction tells us that things are more compli-

cated, and an examination of the $\text{SEX} \times \text{TREAT}$ means reveals that males do particularly well with Treatment 2.

Insert art from p. 411 of previous edition

Insert art from p. 412 of previous edition

17. Below is selected printout from SPSS for Windows 12.0. From the table we can see that none of the effects are significant at the .05 level.

Insert art from p. 413 of previous edition

CHAPTER 5

1.

UNIVARIATE REPEATED MEASURES ANALYSIS						
	TREATMENTS				ROW	
	1	2	3	4	MEANS	
<i>SS</i>	1	5	6	2	5	4.5
	2	3	4	1	6	3.5
	3	3	7	4	10	6.0
	4	6	8	3	3	5.0
	5	4	9	7	8	7.0
	6	5	7	4	9	6.25
	7	2	10	1	2	3.75
	8	4	3	2	5	3.50
	4	6.75	3	6	4.9375	(GRAND MEAN)

$$SS_b = 8[(4 - 4.9375)^2 + (6.75 - 4.9375)^2 + (3 - 4.9375)^2 + (6 - 4.9375)^2] \\ = 72.374$$

$$MS_b = 72.374 / 3 = 24.125$$

$$SS_w = 12 + 39.5 + 28 + 56 \\ \text{sum of squares for Treat 1 Treat 2 Treat 3 Treat 4} \\ SS_w = 135.5$$

SUM OF SQUARES FOR BLOCKS

$$SS_{bl} = 4[(4.5 - 4.9375)^2 + (3.5 - 4.9375)^2 + \cdots + (3.5 - 4.9375)^2] \\ = 51.375$$

$$SS_{res} = 135.5 - 51.375 = 84.125$$

$$MS_{res} = 84.125 / 21 = 4.006$$

$$F = 24.125 / 4.006 = 6.022$$

(a) The critical value at the .05 level, on 3 and 21 degrees of freedom, is 3.07. Therefore, we have a significant overall difference.

(b) Tukey post hoc procedure—The critical value against which each mean difference is to be compared is

$$q_{.05;4,21} \sqrt{MS_{res} / n} = 3.95 \sqrt{4.006 / 8} = 2.795$$

The only mean differences that exceed 2.795 in absolute value are for Groups 2 and 3, and Groups 3 and 4. Thus, these are the only pairs of groups that are significantly different at the .05 level with the Tukey procedure.

3. (a) First, we compute the basic quantities that are to be plugged into the formula:

$$\bar{s}_{ii} = (76.8 + 42.8 + 64)/3 = 61.2$$

This is the average of the diagonal elements.

$$\bar{s} = (76.8 + 53.2 + 69 + 53.2 + \cdots + 47 + 64)/9 = 58$$

This is the average of all the elements in the covariance matrix.

$$\Sigma s_{ij}^2 = 76.8^2 + 53.2^2 + 69^2 + \cdots + 47^2 + 64^2 = 31426.56$$

This is just the sum of all the squared elements in the matrix.

- (b) \bar{s}_i = these are the row averages

$$\bar{s}_1 = (76.8 + 53.2 + 69)/3 = 66.333$$

$$\bar{s}_2 = (53.2 + 42.8 + 47)/3 = 47.667$$

$$\bar{s}_3 = (69 + 47 + 64)/3 = 60$$

$$\hat{\epsilon} = \frac{9(61.2 - 58)^2}{2[31426.56 - 6(10272.21) + 9(3364)]}$$

- (c) Recall that the $\min [??] = 1/(k - 1)$, where k is the number of levels for the repeated measures factor. Since $k = 3$ here, it follows that $\min \epsilon = 1/(3-1) = .50$.

- (d) For the design, there were two groups, with eight subjects per group, and five repeated measures. Thus we have $g = 2$, $n = 8$ and $k = 5$. Therefore,

$$\tilde{\epsilon} = \frac{8(2)(4)(.44629) - 2}{4[2(7) - 4(.44629)]} = \frac{26.56256}{48.85936} = .54365$$

5. Below we present selected printout which gives the F tests for the contrasts and the associated p values:

VARIABLE	HYPOTH. MS	ERROR MS	F	SIG OF F
HELMERT1	24.29999	2.76889	8.77608	.016
HELMERT2	16.53750	1.08639	15.22245	.004
HELMERT3	.00050	.18272	.00274	.959

If overall α is set at .10, then each contrast is being tested at the $.10/3 = .0333$ level of significance. Thus, the first two Helmert contrasts are significant.

The first Helmert contrast is testing whether the control group differs from the remaining 3 groups (the three treatment or drug groups here), while the second Helmert contrast is testing whether the effect of Drug Type I differs from that of the two remaining drugs (which were similar in composition).

CHAPTER 6

1. (a) There definitely does appear to be a linear relationship.
- (c) There is not a pattern in the residuals, which indicates a linear model is appropriate (see Fig. A.3).

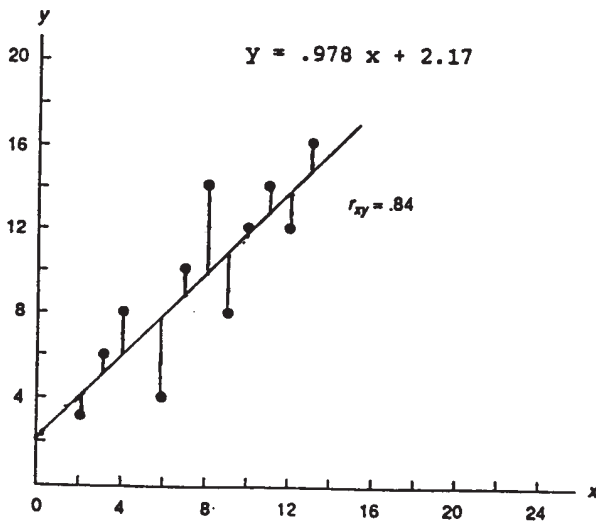


FIGURE A.3

3. (a) If x_1 enters the equation first, it will account for $(.60)^2 \times 100$, or 36% of the variance on y .
 (b) To determine how much variance on y predictor x_1 , will account for if entered second we need to partial out x_2 . Hence we compute the following semi partial correlation:

$$\begin{aligned} r_{y1.2(s)} &= \frac{r_{y1} - r_{y2}r_{12}}{\sqrt{1 - r_{12}^2}} \\ &= \frac{.60 - .50(.80)}{\sqrt{1 - (.8)^2}} = .33 \\ r_{y1.2(s)}^2 &= (.33)^2 = .1089 \end{aligned}$$

Thus, x_1 accounts for about 11 % of the variance if entered second.

- (c) Since x_1 and x_2 are strongly correlated (multicollinearity), when a predictor enters the equation influences greatly how much variance it will account for. Here when x_1 entered first it accounted for 36% of variance, while it only accounted for 11% when entered second.
5. (a) For STEPWISE regression the model selected has SIZE, NEW and NO-BATH as predictors.
 (b) For BACKWARD elimination the same model is selected.

7.

```
TITLE 'USING FIXED FORMAT AND TESTING SET OF
PREDICTORS' .
DATA LIST FIXED/X1 1 X2 2 X3 3-4 X4 5-6 X5 7-8
X6 9-11(2) X7 12-14 X8 15-16.
BEGIN DATA.
DATA LINES
END DATA.
LIST.
REGRESSION VARIABLES = X1 TO X8/
DEPENDENT = X8/
ENTER X1 X2/TEST(X3 X4 X5) /.
```

CHAPTER 7

1. (a) ANCOVA is appropriate since there is a significant linear relationship ($p = .000$) and the homogeneity of regression slopes assumption is tenable ($p = .521$).

(b) We do reject the hypothesis of equal adjusted means since $F = 8.61$ with $p = .001$.

(c) The error terms are related by the equation:

$$MS_w MS_w (1 - r_{xy}^2)$$

3. The error term for the ANCOVA is considerably smaller: 111.72 versus 139.98 for the ANOVA on the difference scores. The regression coefficient from the ANCOVA is .69876. In 5.10 it was stated that, "...whenever, the regression coefficient is not equal to 1, the error term for ANCOVA will be smaller than that for the gain score analysis and hence the ANCOVA will be a more sensitive or powerful analysis of variance assumption since the cell sizes were approximately equal, and ANOVA is known to be robust in this situation."

From 4.6, the relationship between the interaction effect size and the test statistic is given by

$$\hat{f} = \sqrt{(r-1)(c-1)F/N}$$

Therefore, $\hat{f} = \sqrt{(2-1)(2-1)4.49/34} = .3634$

The effect size is fairly large. The n that we would use to enter Cohen's power tables is

$$n = [(N - rc)/(r-1)(c-1) + 1] + 1$$

$$n = [(34 - 4)/(2-1)(2-1) + 1] + 1 = 16$$

Since the degrees of freedom for interaction here is 1, we use Table C.1 and find that power is around .50. Thus, although power was not good in this study, nevertheless significance was found.

5. The fact that the correlation is .61 and that the homogeneity of slopes assumption is tenable means that ANCOVA is appropriate. The grand mean for the study (assuming equal n per group) is 110. Therefore, when the means on the dependent variable are adjusted they will be drawn much closer together, causing a much smaller mean sum of squares between and the loss of significance. The mean of 70 for Group 1 will be adjusted downward (perhaps to a value of 67) while the mean for Group 2 will be adjusted upward (perhaps to a value of 63). Thus, the adjusted means for the 3 groups would be 67, 63, and 65.

Author Index

A

Adelman, H., 230
Agresti, A., 230, 275, 282, 361, 391, 392, 393
Anderson, N. H., 303
Anscombe, F., 228, 229

B

Baker, F., 189
Barcikowski, R., 61, 189
Becker, B., 110
Berry, J., 97
Bloom, B., 59
Bock, R. D., 57, 182, 212, 390
Bosker, R., 323, 324, 343
Bounds, W., 182, 317, 318, 323
Box, G. P., 24, 25, 168, 188, 207, 208, 209, 214
Breen, L., 189
Brown, M. B., 73, 389
Browne, W., 358
Bryan, T., 174, 285, 314, 317
Bryant, J. L., 285, 314, 317
Bryk, A. S., 59, 303, 322, 324, 325, 327, 328, 329, 337, 338, 339, 341, 342, 343
Burststein, L., 321

C

Carlson, J., 144
Chase, C., 21, 160, 161, 329
Cheong, Y. F., 325, 329, 337, 338
Cochran, W. G., 58, 165, 287, 319
Coggeshall, P., 390
Cohen, J., 3, 76, 108, 109, 113, 116, 118, 166, 167, 236, 258
Cohen, P., 76, 258

Collier, R., 189
Congdon, R., 325, 329, 337, 338
Cook, R. D., 47, 227, 240, 246, 261, 262, 263, 264, 266, 273, 277
Copenhaver, M. D., 162
Cormier, W., 182
Cradler, J., 146
Cronbach, L., 111, 124, 150, 172
Crowder, R., 255
Crystal, G., 230

D

Dance, K., 124, 315
Daniels, R., 125, 145
Davenport, J., 59, 323
Davidson, M. L., 189
de Leeuw, J., 323, 324, 326, 342, 343, 358
Delaney, H. D., 353
Dizney, H., 234
Draper, D., 358
Draper, N., 232, 249, 250, 254
Dudycha, A., 263, 265, 273
Dunnett, C. W., 68, 69, 92, 98, 99

E

Elashoff, J., 203, 204, 217, 287, 303

F

Feldt, L., 187, 189, 215, 217
Feshbach, S., 230
Forsythe, A. B., 73
Frane, J., 164

G

Games, P., 41, 73, 74, 77, 93

Geisser, S., 188, 204, 215, 217
 Glass, G., 57, 60, 79, 102, 182
 Glasnapp, D., 222, 304
 Goldstein, H., 358
 Goodwin, D., 146
 Greenhouse, S., 188, 204, 215, 217
 Gromen, L., 234

H

Hagen, E., 307
 Hand, D. J., 141
 Harrington, S., 113
 Hayes, T., 189
 Hays, W., 76, 80, 85
 Hayter, A., 68
 Herzberg, P. A., 239, 254, 258
 Hoaglin, D., 261
 Holland, B. S., 162
 Holm, S., 162, 163
 Hopkins, K., 60, 79, 102, 182
 Howell, J. K., 73, 74, 77, 93
 Hox, J. J., 323, 324, 328, 348, 359
 Huberty, C. J., 68, 238
 Huck, S., 182, 305, 307, 314, 317, 318
 Huitema, B., 287, 293, 304, 308, 310
 Huynh, H., 187, 189, 215, 217
 Hyman, R., 173

J

Jennings, E., 305, 314
 Johnson, P. O., 145
 Johnson, R., 164
 Jones, L. V., 40, 95, 153, 246, 390
 Judd, C., 61, 321, 322

K

Kaiser, M., 189
 Kenny, D., 61, 321, 322
 Keppel, G., 183, 185
 Kerlinger, F., 76
 Keselman, H. J., 189, 197, 201
 Keselman, J. C., 197, 201
 Kirk, R., 76, 318
 Krefl, I., 321, 323, 324, 326, 342, 343, 358

L

Lepine, D., 187
 Levin, J., 97, 98

Light, R., 109
 Lindzey, G., 246, 390
 Littell, R. C., 358
 Longford, N. T., 358
 Lord, F. M., 236, 303

M

Mallows, C. L., 238, 240, 247, 263, 272, 278
 Mandeville, C. K., 189
 Marwit, S., 145
 Maxwell, S. E., 193, 194, 201, 353
 McCabe, G., 250
 McCormick, C., 97
 McLean, R. A., 305, 307, 314
 Mendoza, J., 189
 Miller, G., 97
 Milliken, G. A., 358
 Moore, D., 250
 Morrison, D. F., 239, 240, 241, 244, 250, 259,
 261, 262, 264, 282
 Murray, R. M., 69
 Myers, J., 144, 175, 185, 195, 198, 293, 294,
 295, 297, 298, 301
 Myers, R., 233, 235, 238, 277, 283

N

Neufeld, R., 124
 Neumann, G., 145
 Neyman, J., 285, 287, 300, 308, 311, 314, 315
 Novick, M., 236
 Novince, L., 315
 Nunnally, J., 231, 257

O

O'Brien, R., 189
 O'Grady, K., 76, 79, 96
 Overall, J., 138

P

Park, C., 263, 265, 273
 Paulson, A. S., 285, 314, 317
 Peckham, P., 57
 Pedhazur, E., 76, 144, 304
 Pillimer, D., 109
 Plewis, I., 358
 Poggio, J., 222, 304
 Porter, A., 304
 Presley, M., 97
 Pukulski, J., 172

R

Rasbash, J., 358
 Raudenbush, S. W., 322, 324, 325, 327, 328,
 329, 337, 338, 339, 341, 342, 343, 358
 Reichardt, C. S., 303
 Robey, R., 189
 Rogan, J., 69, 189
 Rogosa, D., 300
 Rosenthal, R., 79
 Rosnow, R., 79
 Rounet, H., 187

S

Sanders, J., 57
 Scariano, S., 59, 323
 Schutz, W., 258, 259
 Serlin, R., 73
 Shiffler, R., 13, 15
 Singer, J. D., 2, 239, 246, 358
 Smith, H., 232, 249, 250, 254
 Smith, S., 40, 95
 Snijders, T., 323, 324, 343
 Snow, R., 79, 111, 124
 Spiegel, D., 138
 Steber, J., 173
 Stein, C., 239, 253, 254, 255, 256, 261, 263,
 265, 272, 273, 277, 321, 358
 Stevens, J. P., 3, 96, 125, 142, 145, 187, 190,
 195, 228, 236, 256, 260

Stoloff, P. H., 189
 Stroup, W. W., 358

T

Taylor, C. C., 141
 Thorndike, R. L., 307
 Timm, N., 144
 Tomarkin, J. J., 73
 Tuckman, B., 173

W

Wagh, M., 95
 Weisberg, H. I., 303
 Weisberg, S., 227, 250, 261
 Welch, B. L., 73, 74
 Well, A., 297
 Welsch, R., 261
 Wichern, D., 164
 Wilkinson, L., 258
 Willett, J. B., 2, 239
 Williamson, F., 230
 Wolfinger, R. D., 358
 Woodhouse, G., 358

Y

Yang, M., 358

Subject Index

A

- Actual alpha, 57, 323
- Analysis of covariance (ANCOVA), 286
 - adjusted means, 289–291
 - alternative analyses, 305–306
 - assumptions, 297, 300
 - choice of covariates, 293
 - computer example with 2 covariates, 312–314
 - computer example, 304–305
 - covariate by treatment interaction, 300
 - homogeneity of regression slopes, 297–300, 306, 308
 - by multiple regression, 297
 - null hypothesis, 294
 - numerical example, 293–294
 - purposes, 287–288
 - reduction of error variance, 288, 292–293
- Analysis of variance (ANOVA) examples, 45
 - assumptions, 56
 - computer example on SAS and SPSS with Tukey procedure, 62
 - computer example with unequal variances and Games-Howell and Tamhane procedures, 77, 93
 - expected mean squares, 53
 - F* test, 51
 - linear model, 55
 - numerical example
 - between group variation, 49–50
 - within group variation, 50–51
- Aptitude by treatment interaction research (ATI), 124
- ASCII file, 39–40

B

- Balanced design, 136
- Bonferroni inequality, 80–81
 - improved Bonferroni type procedure, 162–163

C

- Central Limit Theorem, 57
- Circularity, 187
- Compact disk, 39–40
- Compound symmetry, 187
- Conservative, 58, 68
- Contrast, 71, 82
- Counterbalancing, 185, 213

D

- Dataset editing, 25–28
- Data files, 21, 23
- Dummy coding (group membership), 267, 270
- Dunnett procedure, 68, 92

E

- Effect size
 - factorial ANOVA, 168
 - one way ANOVA, 168
 - t* test, 168
- Empathy model data, 355
- Epsilon
 - Greenhouse–Geisser, 215
 - Huynh–Feldt, 189

Eta squared, 78
 Excel (spreadsheet program), 23
 Expected mean squares, 53

F

Factorial analysis of variance
 advantages, 124
 balanced design, 136
 four way, 160–161
 interpretation of effects, 146
 numerical example for two way, 127–133
 on SAS and SPSS, 153
 three way, 144–157
 two computer examples, 138–142
 unbalanced design, 136–137
 Fixed effects, 169
 Fixed factor, 169–170

H

Harmonic mean, 69
 Heterogeneous variances and unequal group sizes, 73
 Hierarchical Linear Modeling (HLM)
 adding predictors, 340, 348
 data analysis of, 329
 datasets, 330–331
 empathy model data, 355
 estimating parameters, 335–338
 evaluating efficacy, 351
 MDM file, 331–335
 multilevel data, single-level analysis, 323
 multilevel model, formulation of, 325
 two-level example, 329
 two-level model, formulation of, 325–328
 two-level unconditional model, 335
 Higher order designs (see factorial ANOVA)
 HLM6, 329
 HLM software output, 338–339, 344–345, 347–348, 356–357
 Homogeneity of variance assumption statistical tests for, 58

I

Importing datasets, 23
 Independence of observations, 59
 Influential points, 227
 Interaction
 disordinal, 125
 ordinal, 125

Intraclass correlation, 59, 323–324

J

Johnson–Neyman technique, 287, 303, 308

L

Level of significance, 47
 Liberal, 58
 Locus of control, 16
 Lotus 1–2–3 (spreadsheet program), 23

M

Main effects, 128, 130
 Measures of association, 75
 Merging files, 28
 Missing data, 31
 Multiple regression
 ANOVA as a special case of regression analysis, 265
 computer examples, 239
 controlling order with SAS and SPSS, 257
 examples of, 230
 Mallow's C_p , 238
 mathematical maximization procedure, 231
 MAXR (from SAS), 246
 model selection procedures
 multicollinearity, 234
 multiple correlation, 231, 233
 variance inflation factor, 235
 substantive knowledge, 236
 model validation
 data splitting, 239, 252–253
 adjusted R^2 , 254
 Press statistic, 283
 order of predictors, 255
 positive bias of R^2 , 258
 preselection of predictors, 257
 sample size (for a reliable prediction equation), 258
 sequential procedures
 forward selection, 237
 backward selection, 237
 stepwise, 237
 Multivariate analysis of variance (MANOVA), 89

N

Normality, 9, 187
 Notebook computer, 16

O

Omega squared, 76
 Orthogonal comparisons, 83
 Outliers, 12
 detecting, 13
 effect on correlation, 14
 in regression analysis, 260
 Output navigator (SPSS), 31
 Overall alpha, 197

P

p values, 67–68
 Partial correlation, 237
 Partial eta squared, 116
 Planned comparisons, 79–87, 212
 on SAS and SPSS, 87
 test statistic, 84–85
 Platykurtosis, 57
 Power
 a priori determination of sample size, 111
 factors dependent on, 106
 post hoc estimation of, 111
 on SPSS MANOVA, 116
 ways of improving, 115

R

Random factor, 169–170
 Regression
 multiple (see multiple regression)
 simple, 219–225
 Repeated measures analysis
 advantages and disadvantages, 184–185
 single group
 univariate approach, 186
 assumptions, 187
 computer analysis on SAS and SPSS, 190
 one between and one within (trend analysis), 194
 one between and two within, 203–208, 210
 planned comparisons, 212
 SPSS syntax setup for Helmert contrasts, 212
 totally within designs, 209, 211

Repeated measures analysis (*continued*)
 univariate and multivariate approaches
 compared, 189
 Residual plots, 250
 Robust, 57

S

Sample variance, 2
 SAS (Statistical Analysis System)
 analysis of covariance, 295
 control lines for correlations, 19
 dependent samples *t* test, 11
 descriptive statistics, 19
 one between and one within repeated
 measures, 196
 one between and two within repeated
 measures, 204
 one way ANOVA, 63
 planned comparisons, 87
 single group repeated measures, 190
 t test, 19
 three way ANOVA, 154
 two way ANOVA, 137
 SAS (selected printouts)
 ANCOVA, 295–296
 MAXR regression, 248
 one between and two within repeated
 measures, 206
 one way ANOVA, 64
 planned comparisons, 91
 single group repeated measures, 192
 two way (disproportional cell *n*), 143
 two way (equal cell *n*) ANOVA, 139–140
 Scheffé procedure, 71
 Split file, 28
 SPSS (Statistical Package for the Social
 Sciences)
 analysis of covariance (Sesame Street data),
 305
 analysis of covariance via regression, 298
 analysis of covariance with 2 covariates, 312
 control lines for correlations, 22
 control lines for univariate nested design, 354
 dependent samples *t* test, 22
 descriptive statistics, 22
 Helmert contrasts in single group design, 212
 one between and one within repeated
 measures, 196
 single group repeated measures, 190

SPSS (*continued*)*t* test, 22two way ANOVA (equal cell *n*), 137

SPSS screens

GLM for disproportional two way ANOVA,
143GLM options screen for obtaining marginal
means, 159one between and two within repeated
measures, 210–211

one way ANOVA and Tukey procedure, 65

planned comparisons, 90

SPSS for Windows (selected printouts)

analysis of covariance (Sesame Street data),
306

ANCOVA with two covariates, 313

one between and one within repeated
measures, 198–200

one way ANOVA, 66

planned comparisons, 91

power analysis (*t* test and one way ANOVA),
117

stepwise and backward selection, 241–243

three way ANOVA, 155–157

three-level nested design, 355

two way ANOVA (equal cell size), 136

two way ANOVA (unequal cell size—
computer example), 144

Sphericity, 187–193

Summation notation, 5–7

T

t tests

independent samples, 8–10

dependent samples, 11–12

Tail probability, 67–68

Tau matrix, 345

Three way interaction effect, 148–150

TI-30Xa calculator, 12, 44, 94

Trend analysis, 194

Tukey procedure, 69–70

Type I error, 8, 105

Type II error, 105

Type I sum of squares (sequential), 141

Type III sum of squares (unique sum of
squares), 141

U

Uniformity, 187

W

Welch test statistic, 73

Windows 12.0 (SPSS), 20–21

Praise for the Second Edition:

"Stevens does an excellent job showing students how to use and read computer output...A strength of this book is the author's very clear explanation of power analysis...The author's presentation style is very readable and easy to follow."

— *The American Statistician*

"...Extremely useful to those conducting educational or other similar research."

— *The Statistician*

"The greatest strength is the accessibility of the material...readers are exposed to differences in approaches and philosophy for data analyses and are given enough information to make informed decisions for themselves...The illustrations and screen dumps are a strong feature."

— *Dale E. Berger, PhD, Dept. of Psychology, Claremont Graduate University*

"I have not found a textbook that manages the link between intermediate statistical concepts and their application better than Professor Stevens' *Intermediate Statistics*...The chapter on power analysis is one-of-a-kind...the conceptual and practical approach...and the emphasis on assumptions of the statistical tests...separates it from others...this is one of the only textbooks to cover Multiple Regression and ANOVA from their own unique perspectives."

— *Gordon P. Brooks, PhD, Dept. of Educational Studies, Ohio University*

"I am currently using Stevens 2/e for my Advanced Statistics course for first year graduate students...my class emphasizes the connections between ANOVA and regression...this was one of the only books that covered both topics on a level accessible for clinical students...I will definitely use the 3rd Edition."

— *Michael Milburn, PhD, Dept. of Psychology, University of Massachusetts-Boston*

James Stevens' best-selling text, *Intermediate Statistics*, is written for those who use, rather than develop, statistical techniques. Dr. Stevens focuses on a conceptual understanding of the material rather than on proving the results. SAS and SPSS are an integral part of each chapter. Definitional formulas are used on small data sets to provide conceptual insight into what is being measured.

The assumptions underlying each analysis are emphasized and the reader is shown how to test the critical assumptions using SPSS or SAS. Printouts with annotations from SAS or SPSS show how to process the data for each analysis. The annotations highlight what the numbers mean and how to interpret the results. Numerical, conceptual, and computer exercises enhance understanding. Answers are provided for half of the exercises.

The book offers comprehensive coverage of one-way, power, and factorial analysis of variance, repeated measures analysis, simple and multiple regression, analysis of covariance, and HLM. Power analysis is an integral part of the book. A computer example of real data integrates many of the concepts. Highlights of the Third Edition include:

- A new chapter on hierarchical linear modeling using HLM6.
- A CD containing all of the book's data sets.
- New coverage of how to cross validate multiple regression results with SPSS and a new section on model selection (Ch. 6).
- More exercises in each chapter.

Intended for intermediate statistics or statistics II courses taught in departments of psychology, education, business, and other social and behavioral sciences, a prerequisite of introductory statistics is required. An *Instructor's Solutions CD* is available to adopters.
